Article

Social Psychological and Personality Science
2024, Vol. 15(7) 812–823

DOI: 10.1177/19485506241256400
journals.sagepub.com/home/spp
Sage

Echoes of Culture: Relationships of Implicit and Explicit Attitudes With Contemporary English, Historical English, and 53 Non-English Languages

Tessa E. S. Charlesworth[1], Kirsten Morehouse[2], Vaibhav Rouduri[2], and William Cunningham[3,4]

Abstract

Attitudes are intertwined with culture and language. But to what extent? Emerging perspectives in attitude research suggest that cultural representations in language are more related to implicitly measured (vs. explicitly measured) attitudes, and that such relationships persist across history and diverse languages. We offer a comprehensive test of these ideas by correlating (a) attitudes toward 55 topics (e.g., *Rich/Poor, Dogs/Cats, Love/Money*) from ~100,000 U.S. English-speaking participants with (b) representations of those same topics in word embeddings from contemporary English text, 200 years of English books, and 53 non-English languages. Strong and robust relationships emerged between representations in contemporary English and implicitly but not explicitly measured attitudes. Moreover, strong correlations with implicitly measured attitudes persisted across 200 years of books, and most non-English languages. Results provide new insights into the nature of implicitly measured attitudes and how they are intertwined with cultural representations that are relatively hidden in patterns of language across time and place.

Keywords

attitudes, culture, implicit attitudes, language, word embeddings

> In the same way that groundwater can be contaminated with environmental pollutants, our minds. . . appear to passively absorb cultural biases that may result in prejudices. . .
>
> —Olson and Kendrick (2011, p. 119)

Culture exerts a powerful force over how we think, feel, and act (Mohr et al., 2019). Yet culture is notoriously hard to define, let alone measure. Early perspectives even warned that culture was like "an amorphous, indescribable mist which swirls around society members" (Fine, 1979, p. 733). Today, however, with advances in natural language processing (NLP) and availability of massive text corpora, we are newly equipped for large-scale, quantitative tests of culture as transmitted through language. That is, by measuring repeated associations of concepts in language (e.g., repeatedly associating bad-insects/good-flowers), we can gain insight into *cultural representations*, defined as concept associations uncovered in large-scale language shared across many people (Durkheim, 1974; Moscovici, 1994).

Quantifying cultural representations is crucial for advancing understanding on the nature of attitudes and especially the similarities or differences between attitudes captured in relatively implicit versus explicit measures (Greenwald & Banaji, 1995). Originally, both implicitly and explicitly measured attitudes[1] were hypothesized to reflect some combination of both cultural inputs and personal values (Banaji, 2001), though the relative influence of these factors remained debated (Karpinski & Hilton, 2001; Nosek & Hansen, 2008). From among these debates, the "Bias of Crowds" perspective emerged to argue that implicitly measured attitudes mostly reflect culture (Payne et al., 2017), with perhaps a small contribution from personal values. In contrast, explicitly measured attitudes are filtered through individual values, ideologies, and choices (Cunningham et al., 2007; van Bavel et al., 2012) and thus less directly reflecting culture.

[1]Northwestern University, Evanston, IL, USA
[2]Harvard University, Cambridge, MA, USA
[3]University of Toronto, Ontario, Canada
[4]Vector Institute, Toronto, Ontario, Canada

Corresponding Author:
Tessa E. S. Charlesworth, Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA.
Email: tessa.charlesworth@kellogg.northwestern.edu

To support this emerging perspective on the nature of attitudes, a key test would be to show strong links between cultural representations in language and implicitly measured attitudes, but weak links between cultural representations and explicitly measured attitudes. In this vein, past work has already provided suggestive evidence for relationships between human attitudes and distributed patterns in large language corpora (reviewed in Charlesworth & Banaji, 2022b). For example, Caliskan and colleagues (2016) documented eight biased associations in internet text that were also observed on measures of human attitudes. They concluded that cultural representations extracted from large-scale text must therefore be reflections of implicitly measured attitudes, a conclusion that has since been advocated by many others as well (Bhatia & Walasek, 2023; Bolukbasi et al., 2016; Charlesworth & Banaji, 2022b; Charlesworth et al., 2023; Hauser & Schwarz, 2022; Lewis & Lupyan, 2020).

These interpretations largely stem from the observation that biases were present in both language patterns and implicitly measured human attitudes. For example, an anti-Old/pro-Young association was observed in internet text and in humans' implicitly measured attitudes (Caliskan et al., 2016). However, anti-Old/pro-Young associations are also observed on explicitly measured attitudes (Charlesworth & Banaji, 2022a). Thus, the mere presence of biased associations in both language and implicitly measured attitudes does not determine whether, and to what extent, cultural representations are *relatively* more related to implicitly or explicitly measured attitudes which, again, is the key test of recent perspectives on the nature of attitudes. Although recent work sought more direct tests of such relative relationships, the work is still limited in considering only one topic (e.g., only gender stereotypes; Lewis & Lupyan, 2020), only within individual participants (e.g., how words prime individuals' attitudes; Hauser & Schwarz, 2022), or only within one cultural frame (e.g., contemporary English text; Bhatia & Walasek, 2023). The current work offers a first comprehensive direct test of the relative strength of association between cultural representations in language and implicitly versus explicitly measured attitudes across a set of 55 diverse attitude topics examined on aggregate across ˜100,000 participants and billions of words of text.

Critically, the current case also goes beyond past approaches to test two further hypotheses, elaborated below. First, we test whether implicitly measured attitudes will be tied to cultural representations even in historical language because of long-term cultural transmission. After all, one defining aspect of culture is that it is transmitted across generations, with some cultural values showing remarkable persistence (Bruner, 1956). We expect that implicitly measured attitudes reflect these relatively persistent cultural representations. Thus, contemporary implicitly measured attitudes will still be linked with language from long ago (e.g., the 1800s). Explicitly measured attitudes, however, are primarily filtered through personal values, which will update alongside changes in acceptability of expressing certain prejudices (Crandall et al., 2002; Devine et al., 2002; Marsden et al., 2020). Thus, if any relationship between explicitly measured attitudes and cultural representations exists, we expect it to be only with representations in contemporary language (from the 2000s) which express the most similar ideologies.

Second, we also consider the prediction that implicitly measured attitudes will be tied to cultural representations from non-English languages because of widespread cultural sharing. Indeed, a second distinguishing aspect of culture is that it can be shared across languages and places, such as through "horizontal borrowing" between neighboring languages and countries (Greenhill et al., 2009). Cultural representations that are shared across English and non-English languages may therefore also be related to the implicitly measured attitudes from English-speaking U.S. participants. This would mean, for example, that cultural representations in Spanish, French, or even Urdu and Telugu, might reflect similar representations (e.g., a preference for Rich vs. Poor or Love vs. Money) that, in turn, are intertwined with implicitly measured attitudes from English-speaking U.S. participants. This result would further underscore the widespread, shared nature of implicitly measured attitudes.

This link between implicitly measured attitudes and cultural representations is expected to be widespread across most non-English languages. Yet it is possible that relationships will be strongest with non-English languages that are relatively closer to English, whether through shared linguistic roots (e.g., Indo-European languages) or geographic proximity (e.g., languages spoken in France or Spain versus Madagascar or Sri Lanka; Greenhill et al., 2009). In contrast, explicitly measured attitudes from English-speaking U.S. participants are not expected to be related to cultural representations from other languages because these non-English cultures have different norms about what is acceptable to express. In summary, through contemporary, historical, and multi-language analyses, the current work tests emerging ideas about the nature of implicitly versus explicitly measured attitudes and how they are intertwined with culture in language.

Method

Open Practices

All data and code are openly available through the project's OSF page: https://osf.io/9kg5h/. All analyses, data collection, and preparation procedures for contemporary English language analyses were formally preregistered at https://osf.io/d63pt. Additional analyses with historical data and non-English languages were not preregistered.

Table 1. *Contemporary English-Language Pretrained Embedding Datasets*

Language dataset name	Dataset codename	Word tokens (total size)	Unique words (vocabulary)
GloVe Common Crawl	gcc1	840 billion	2,196,017
GloVe Common Crawl-2	gcc2	42 billion	1,917,494
GloVe Wikipedia	gwiki	6 billion	400,000
GloVe Twitter	gtwit	27 billion	1,193,517
word2vec Google News	wnews	100 billion	692,000
word2vec Google Books (1990)	wbooks	~100 billion	71,097
fastText Wikipedia	ftwiki	16 billion	999,994
fastText Common Crawl	ftcc	600 billion	2,000,000
PPMI New York Times (2015)	pnyt	94 million	20,936

Note. See the Supplementary Appendix for additional details and sources are provided in for all embedding datasets. The *word2vec* Google Books dataset comprises ~500 billion word tokens across all 200 years; however, the estimate for the last decade of text alone is ~100 billion word tokens, which corresponds to the 71,097 unique words. Similarly, the *PPMI* NYT dataset comprises 94 million word tokens across all 26 years, with a vocabulary of 20,936 unique words that are available across all years (see training details in the Supplementary Appendix). The *fastText* Wikipedia dataset also includes text data from UMBC WebBase corpus and statmt.org news crawl corpus.

Data Sources

Attitude Data: Attitudes, Identities, and Individual Difference Dataset. We began with a large dataset of humans' implicitly and explicitly measured attitudes from the *Attitudes, Identities, and Individual Differences* (AIID) project from the work by Hussey and colleagues (2012). English-speaking U.S. participants were randomly assigned to complete attitude and belief measures for 95 topics. Based on inclusion criteria that allow us to compute accurate measures of these topics in language (see below), we use 55 topics, with a final sample of 96,605 participants (Supplementary Appendix for demographics).

Implicitly measured attitudes were operationalized with D-score estimates from the Implicit Association Test (IAT; Greenwald et al., 1998). Explicitly measured attitudes were operationalized with self-reported seven-point Likert-type scales, from −3 (e.g., *strongly prefer Gay over Straight people*) to + 3 (*strongly prefer Straight over Gay people*), with 0 reflecting no explicit preference index explicitly measured attitudes. Additional explicit beliefs were also collected (e.g., about attitude strength, certainty, and cultural pressures). We use these additional explicit measures in preregistered exploratory analyses below. Further details on the AIID data are at https://osf.io/pcjwf/.

Text Data: Pretrained Word Embeddings. For the primary contemporary English analyses, we use nine pretrained datasets of word embeddings—vector representations of word meaning that have already been trained and validated from massive corpora of naturalistic text (Table 1).

For historical analyses, we used all available decades from the *word2vec* Google Books dataset, with 20 decades covering 1,800–2,000 (Hamilton et al., 2016). For non-English languages, we used *fastText* embeddings trained on Wikipedia and Common Crawl text from 53 non-English languages (Grave et al., 2018). These 53 languages were chosen because they were among the most commonly spoken languages around the world and had available pretrained embeddings. To translate the ˜35,000 word stimuli, we used GPT3.5 as an automated dictionary (i.e., asking ChatGPT to translate all stimuli into French, Spanish, and so on). Fluent speakers checked a subset of translations for accuracy and provided nearly identical translations when performed by hand, indicating that the automated translations provided accurate word stimuli.

Extracting Cultural Representations From Text: Selecting Topics and Word Stimuli. From the possible list of 95 topics in the AIID dataset, we first determined which of these topics could be accurately computed in text. Because we used pretrained static word embeddings, words are represented as one-word phrases that collapse all meanings to one embedding (e.g., "bank" has only one embedding). As such, we excluded all attitude topics requiring two-word stimuli (e.g., *Bill Clinton*) or with high polysemy (e.g., *coke*, which could refer to soda or drugs). Inclusions were determined by three independent coders, resulting in 55 final topics (Supplementary Appendix for rating study). Stimuli lists were then created using as many original stimuli from AIID as possible. If the original stimuli were images (7 of the 55 final topics), or highly polysemous (17 of the 55 final topics), words were added using online thesaurus searches (Table S2 in Supplementary Appendix).

Extracting Cultural Representations From Text: Four Methods of Extraction. Cultural representations were extracted from language using four approaches,[2] each elaborated below. All methods relied on computing cosine similarities between word embedding vectors. Cosine similarities are essentially correlations between the vectors of numbers used to represent two words (e.g., the correlation between the vectors to represent *good* and *flowers*). Higher cosine similarities indicate more semantic overlap because the vectors are close together in the word embedding space.

Word Embeddings Association Test. The first two methods use the Word Embeddings Association Test (WEAT), which is computed much like an IAT D-score and comparable to Cohen's *d* effect sizes. WEAT begins by taking four lists of words—Group-A, Group-B, Positive, and Negative (Caliskan et al., 2016). With these words, we compute the relative average cosine similarities between the Group word vectors and the Positive/Negative word vectors, and normalize by the pooled standard deviation across all pairs of cosine similarities. To illustrate, for a WEAT of Straight (e.g., represented by words *heterosexual, hetero, straight*) versus Gay (e.g., *homosexual, gay, lesbian*), we first take the average cosine similarities between Straight words and all positive words, yielding a Straight-positive association. Second, we repeat this for negative words, yielding a Straight-negative association. Third, we take the difference between these associations to get a relative Straight-positive/negative association. Fourth, we repeat these three steps for Gay words, yielding a relative Gay-positive/negative association. Fifth, we subtract the Gay-positive/negative from the Straight-positive/negative to yield the relative positive/negative differences between groups, and normalize by the pooled standard deviation across all pairs of cosine similarities between all words. WEAT D-short is computed using positive/negative words that are identical to those used in the AIID dataset (˜7 positive, 7 negative words); WEAT D-long is computed using the top 100 positive/negative words (from humans' ratings provided in the work by Warriner et al., 2013).

Mean Average Cosine Similarity Valence. Valence from Mean Average Cosine Similarity (MAC) is computed by taking two lists of words, Group-A and Group-B, and then identifying the top-*N* words that have the highest relative cosine similarity to Group-A over Group-B (Charlesworth et al., 2022). Next, each of the top-*N* words is replaced with a corresponding valence score from human ratings (Warriner et al., 2013) from −4 (*very negative*) to + 4 (*very positive*). The average across the top-*N* words' valence scores is the valence score of Group-AvB. We repeat this process to compute the Group-BvA valence score and take the difference between Group-AvB vs. Group-BvA, resulting in a relative valence score analogous to the WEAT and IAT.

To illustrate, for MAC valence of Straight versus Gay, we first find the top-10 words with the highest cosine similarity to Straight words and the lowest cosine similarity to Gay words, resulting in words such as {*conventional, stable, unaffected*}, which have an average valence of + 0.94. We then repeat the process for the top-10 words associated to Gay versus Straight, resulting in words such as {*stupid, serious, worried*}, with an average valence of −0.47. Finally, we take the difference between these two average valence scores (0.94 − [−0.47]) to yield a relative MAC score of + 1.41. MAC-short is computed by pulling the top-10 words; MAC-long is computed by pulling the top-50 words.

Notes on Sample Size, Power, and Analytic Strategy. The final sample of contemporary language is N = 1,980, derived from 55 topics with 36 extracted language representations per topic (4 extraction methods by 9 text data sources). Historical language analyses use N = 4,400 observations derived from 55 topics with 80 extracted language representations per topic (4 extraction methods by 20 decades). Non-English language analyses use N = 2,915 observations derived from 55 topics with 53 extracted non-English representations.

Although each observed estimate is highly precise because it is based on large data from human participants and text sources, the nested sample sizes nevertheless set upper limits on power. As such, we perform all analyses using Bayesian inference with preregistered Regions Of Practical Equivalence (ROPE; Kruschke, 2018). This modeling approach is less vulnerable to power-related concerns of Type II errors from null hypothesis significance testing. Instead, the percentage of model posteriors falling inside the ROPE can be used to directly quantify the amount of evidence in favor of a hypothesized result (i.e., a meaningful relationship between language representations and attitudes). For inference, we preregistered that > 95% of the posterior falling outside the ROPE (set at $0.1 \times SD_y$) would provide strong evidence in favor of the alternative hypothesis (i.e., that language is related to attitudes). If > 95% of the posterior falls inside the ROPE, we have strong evidence in favor of the null. Any percentages below 95% that fall inside/outside the ROPE are interpreted as inconclusive evidence.

Bayesian linear regressions were fit using the *stan_glm()* function from the *rstanarm* package in *R* (Goodrich et al., 2020). All models converged with default parameter settings: weakly informative priors were automatically adjusted by *rstanarm* (in this case, the prior for the coefficient was centered at 0, with a scale of 1.8), with default MCMC (Markov chain Monte Carlo) sampling settings (four chains; 2000 iterations, with 1000 iterations discarded from burn-in).

Results

Are Implicitly or Explicitly Measured Attitudes More Related to Cultural Representations?

Results showed strong evidence in favor of a relationship between cultural representations and implicitly measured attitudes, $Med_{posterior}$ = 0.45, *95% Credible Interval* ($CI_{posterior}$) = [0.30, 0.61], with 100% of the posterior falling outside the ROPE (Figure 1A). This moderate relationship was also robust across 32 of 36 model specifications (Figure 2). Implicitly measured attitudes and cultural representations thus appear to be intertwined across nearly all large-scale language dataset in use today. In contrast, we found only weak evidence for a relationship between cultural representations and explicitly measured attitudes, $Med_{posterior}$ = 0.15, $CI_{posterior}$ = [−0.05, 0.34], with 68% of

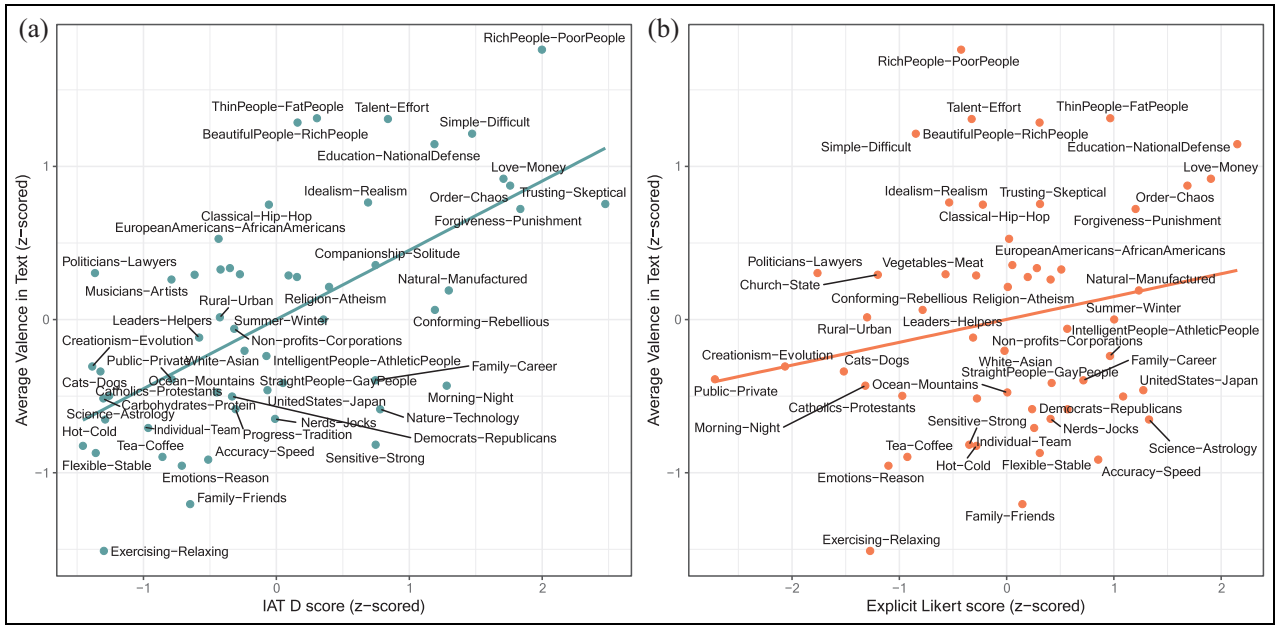

Figure 1. *Relationship Between Attitudes and Cultural Representations Drawn From Contemporary Pretrained English-Language Word Embedding Models. Panel A Depicts the Relationship With Implicitly Measured Attitudes; Panel B Depicts the Relationship With Explicitly Measured Attitudes*
Note. Higher scores on the *x*-axes (to the right of the plot) indicate that human participants held more positive relative attitudes toward the first concept (e.g., "Rich People") over the second concept (e.g., "Poor people"). Similarly, higher scores on the *y*-axes (to the top of the plot) indicate that the text had more positive relative associations with the first concept over the second concept.

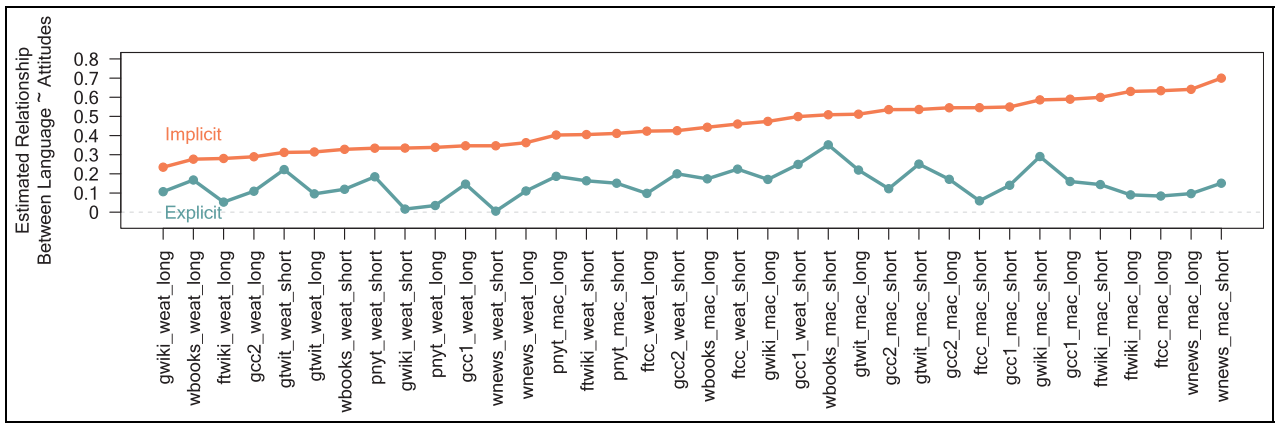

Figure 2. *Relationship Between Attitudes and Cultural Representations From Contemporary English-Language Text Across 36 Model Specifications*
Note. Model estimates are derived from the four methods of extracting valence from text for each of the nine commonly pretrained embedding models listed in Table 1. Bayesian estimates are ordered from the smallest (on the left of the plot) to the largest (on the right of the plot) based on the strength of relationship with implicitly measured attitudes. Name codes of the embedding models are provided in Table 1.

posterior falling outside the ROPE, indicating more ambiguous evidence (Figure 1B). Only 2 of 36 model specifications had > 95% of their posterior outside the ROPE (Figure 2).

Entering both aggregate implicitly and explicitly measured attitudes simultaneously (as preregistered), we found that the relationship between cultural representations and implicitly measured attitudes continued to be meaningful (with 100% of the posterior falling outside the ROPE), but the relationship with explicitly measured attitudes weakened, such that only 26% of the posterior for the explicit attitudes' estimate fell outside the ROPE.

Moderation by Attitude Topic Domain. The diverse 55 topics enabled a secondary (preregistered) exploratory analysis of

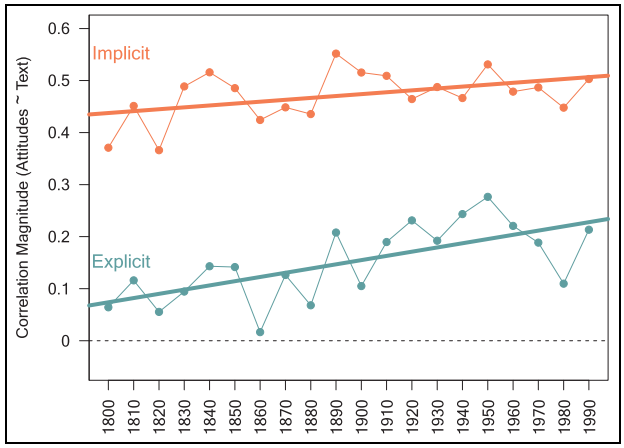

Figure 3. *Relationships Between Contemporary Attitudes and Cultural Representations in Historical Text Across 20 Decades*
Note. Y-axis indicates magnitude of correlation between attitudes and language (averaged across four methods of extraction from language). X-axis indicates the decade of the text data source. Orange lines indicate correlations with implicitly measured attitudes; blue lines indicate correlations with explicitly measured attitudes. Thin lines indicate raw correlation estimates; thick straight lines indicate fitted Bayesian slope.

whether some domains of attitude topics may have stronger relationships between cultural representations and attitudes. Perhaps explicitly measured attitudes are distinct from cultural representations primarily for socially sensitive topic domains (i.e., social groups) where specific rules govern the expressions of prejudice; yet relationships could still emerge for less-sensitive topic domains (e.g., concrete objects, abstract concepts). However, as reported in the Supplementary Appendix, we found no meaningful moderation by domain, indicating a robust absence of relationships between explicitly measured attitudes and cultural representations. Similarly, we found no moderation by domain for implicitly measured attitudes and cultural representations, indicating a robustly strong relationship, regardless of domain.

Additional Explicitly Measured Beliefs. Finally, we performed a preregistered exploration of relationships between cultural representations and an additional 45 explicitly measured beliefs (e.g., "there is cultural pressure to think positive things about Straight people"; Supplementary Appendix). Only 11 of these explicitly measured beliefs passed preregistered thresholds for meaningful relationships. These 11 meaningful relationships generally emerged with perceptions of cultural beliefs, such as "how does society evaluate Straight versus Gay people," which had a relationship of $Med_{posterior}$ = 0.33, [0.16, 0.52], 100% outside ROPE. Crucially, however, even these few relationships seemed to be explained by the relationship with implicitly measured attitudes. That is, adding both IAT D-scores and cultural perceptions weakened relationships with cultural perceptions, $Med_{posterior}$ = 0.19, [0.03, 0.36], 90% of posterior outside ROPE, but the IAT D-scores continued to be strongly related to cultural representations, $Med_{posterior}$ = 0.38, [0.21, 0.54], with 100% of the posterior outside ROPE. We return to this important finding in the general discussion. No other explicitly measured personal beliefs (e.g., certainty, polarity) were meaningfully related to language representations (Supplementary Appendix).

Relationship of Attitudes and Cultural Representations Across Time

We next test the prediction that implicitly measured attitudes reflect cultural representations that are relatively persistent over time. Repeating analyses with 200 years of historical book text revealed relationships between implicitly measured attitudes and cultural representations even in the earliest decade (Figure 3). Indeed, the timeseries of 20 decade-wise correlations between implicitly measured attitudes and cultural representations showed no meaningful change in correlations, $Med_{posterior}$ = 0.0036, [0.0008, 0.0064], 0% outside ROPE. Thus, implicitly measured attitudes of contemporary U.S. English-speaking participants appear to reflect cultural representations transmitted from at least 200 years ago.

We again found no meaningful relationships with explicitly measured attitudes and cultural representations in historical text, and no evidence of change in the relationship, $Med_{posterior}$ = 0.0081, [0.0061, 0.010], 0% outside ROPE (Figure 3). Nevertheless, the slope for explicitly measured attitudes was, descriptively, 2x larger than that for implicitly measured attitudes. This may suggest that, to the extent that explicitly measured attitudes and cultural representations have any relationship, it is more likely when both attitudes and text reflect similar contemporary values.

Relationship of Attitudes and Cultural Representations Across Languages

Finally, we test whether implicitly measured attitudes reflect cultural representations that are shared across diverse (non-English) languages. Using the average score from 53 non-English languages, we found a meaningful relationship with implicitly measured attitudes, $Med_{posterior}$ = 0.31, [0.27, 0.35], 100% outside ROPE (Figure 4A). In addition, directly comparing the strength of relationships between English and non-English languages in an interaction model showed no meaningful interaction, $Med_{posterior}$ = −0.14, [−0.34, 0.06], 66% outside ROPE, implying that relationships were similar in strength for English and non-English languages (Supplementary Appendix for further details). Once again, no meaningful relationship appeared between cultural representations in non-English text and explicitly measured attitudes, $Med_{posterior}$ = 0.08, [0.04,

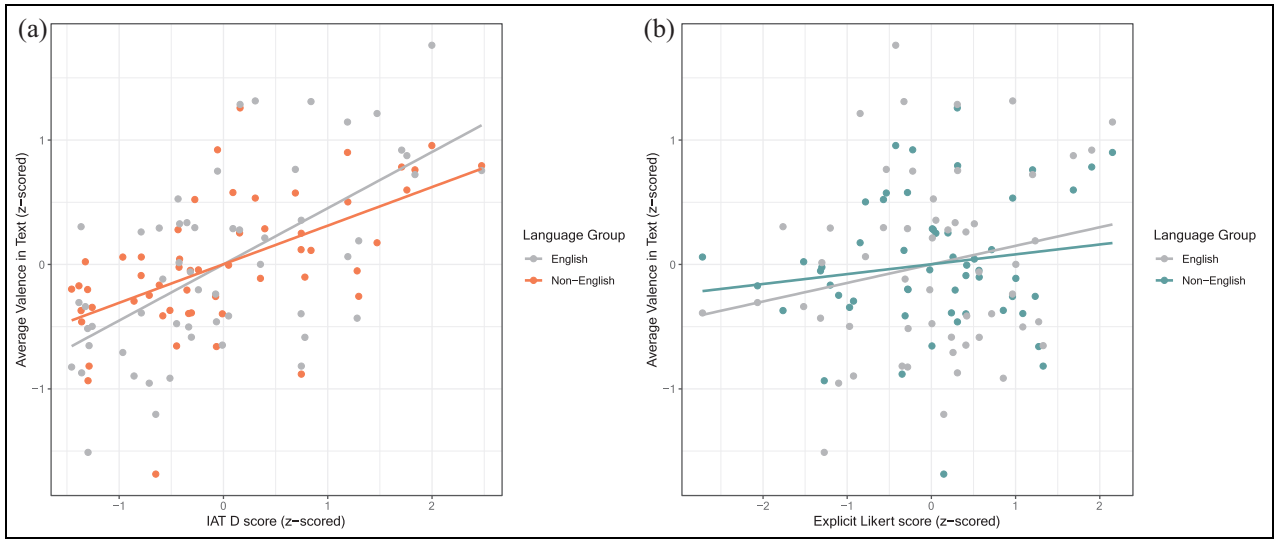

Figure 4. *Relationships Between Attitudes and Cultural Representations in English Versus Non-English Text*
Note. Panel A indicates the relationships with implicitly measured attitudes; Panel B indicates the relationships with explicitly measured attitudes. Y-axis indicates magnitude of the estimated language effect from either English (gray) or non-English (orange or blue text). Individual dots indicate the 55 individual attitude topics. Gray line indicates the estimated relationship between language and attitudes for English text; colored lines (orange or blue) indicate the estimated relationship between language and attitudes for non-English text.

0.12], 15% outside ROPE, with no meaningful interaction, $Med_{posterior} = -0.07$, [−0.31, 0.17], 56% outside ROPE (Figure 4B).

These 53 non-English languages vary considerably in their closeness to English. For instance, languages such as Spanish and French are spoken in countries that are geographically close to the United States and share Indo-European linguistic roots. Other languages such as Malagasy and Malay are spoken in countries much further from the United States and belong to the Austronesian language family. We explore such variation by fitting individual language models (i.e., separate Bayesian regression for Spanish, French, etc.) and examining possible moderators of relationships across non-English languages.

For implicitly measured attitudes, 55% of languages (29 out of 53) showed meaningful links to cultural representations (Figure 5A; full table in Supplementary Appendix), suggesting the two are intertwined even across (most) linguistic boundaries. For instance, the strong positive evaluations in English of Rich versus Poor, Thin versus Fat, or even Love versus Money are also among the strongest associations in Spanish, Hungarian, Telugu, and many others, and such consistent representations are also captured in implicitly measured attitudes. In contrast, explicitly measured attitudes showed no meaningful relation to language estimates in *any* of the 53 languages (Figure 5B).

We next test two moderators: (1) language family, a binary variable (Indo-European versus non-Indo-European languages); and (2) geographic proximity, a continuous variable of the miles between the United States and the closest point of the primary country in which the language is spoken (e.g., closeness to Spain for Spanish).[3] Although descriptive results were in the expected direction, with geographically more distant and non-Indo-European languages showing weaker relationships, neither variable was a meaningful moderator: language families, $Med_{posterior} = -0.07$, [−0.15, 0.00], 22% outside ROPE; geographic proximity, $Med_{posterior} = -0.03$, [−0.07, 0.00], 0% outside ROPE (Figure 5). Other moderators (e.g., rates of immigration between the primary speaking country and the United States, or other indicators of cultural similarity) may more meaningfully explain which languages show stronger links to U.S. English-speakers' implicitly measured attitudes.

General Discussion

Attitudes are an "indispensable construct" in social psychology (Allport, 1954). Yet the nature of this construct seems continuously open to debate and interpretation (Albarracín et al., 2005). These debates seem especially vocal when discussing the similarities and differences of implicitly versus explicitly measured attitudes (e.g., Kurdi et al., 2021; Schimmack, 2021). Among these debates, the Bias of Crowds perspective (Payne et al., 2017) inspired a new suggestion that implicitly measured attitudes are particularly and strongly tied to cultural representations, whereas explicitly measured attitudes are more filtered through personal values. Testing such ideas directly has proved challenging because of the complexities inherent in quantifying cultural representations at scale (Mohr et al., 2019). Here, equipped with NLP, we measured cultural representations in large language corpora to provide a comprehensive analysis of (a) the relative relationship between cultural representations and implicitly versus

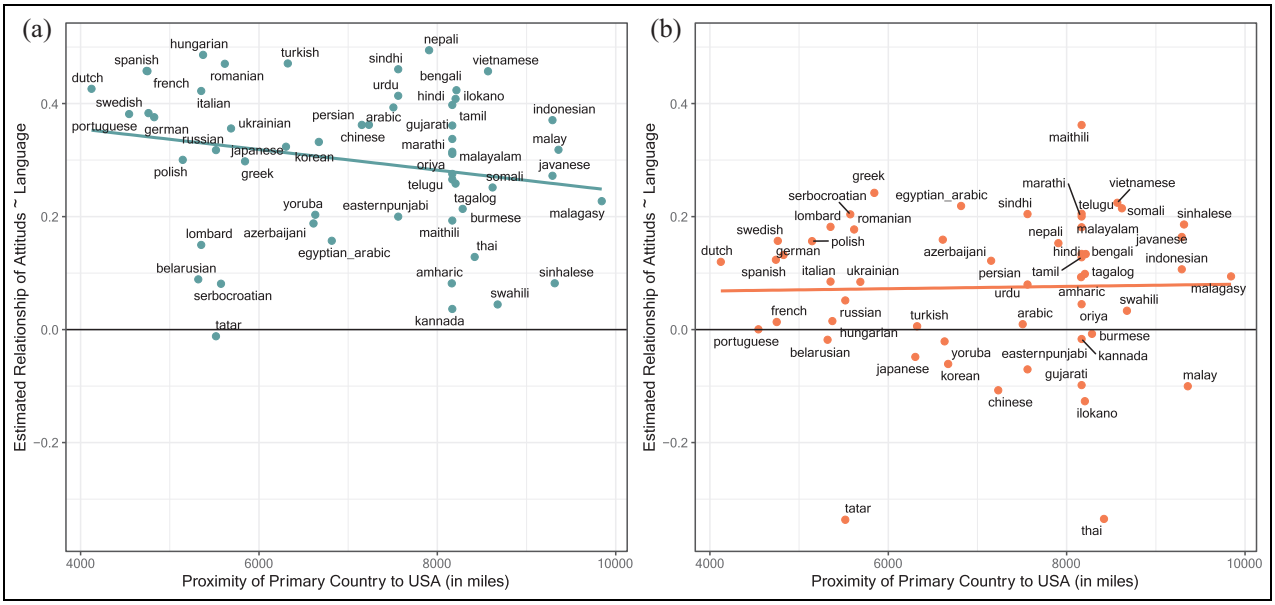

Figure 5. *Relationships Between Attitudes and Cultural Representations in Non-English Languages, Predicted by Geographic Proximity of Language Speakers*

Note. Panel A Indicates the Relationships for Implicitly Measured Attitudes; Panel B Indicates the Relationships for Explicitly Measured Attitudes. X-axis indicates the geographic proximity between the United States and the language speakers' primary country (e.g., Spain for Spanish speakers); higher values indicate more distance from the United States. Y-axis indicates magnitude of the estimated slope between attitudes and cultural representations in language.

explicitly measured attitudes, as well (b) the persistence of such relationships across history, and (c) the spread of relationships across languages. In the process, we arrived at a more nuanced understanding of the nature of implicitly measured attitudes and their unique intertwining with cultural representations in language.

Robust Relationships Between Implicitly Measured Attitudes and Cultural Representations

The clearest result is that implicitly measured attitudes were strongly and robustly linked to cultural representations in large-scale text. Most notably, implicitly measured attitudes correlated with cultural representations in contemporary English text regardless of the underlying text corpus (e.g., *Wikipedia* vs. internet text vs. books). Implicitly measured attitudes are thus not only tied to cultural representations revealed in relatively spontaneous texts (e.g., *Twitter* and internet text) but also representations in more controlled sources (e.g., edited books). Implicitly measured attitudes may therefore relate to the consistent and shared cultural representations (e.g., consistent positive representations of Rich over Poor or Love over Money). Explicitly measured attitudes may be more related to idiosyncratic representations that vary across contexts.

Skeptics may wonder whether the robust link between implicitly measured attitudes and cultural representations is simply because implicitly measured attitudes are, themselves, a proxy for participants' explicit knowledge of culture (Karpinski & Hilton, 2001) and awareness of cultural stereotypes (Devine, 1989; Katz & Braly, 1933). This interpretation is not supported by the current data. Implicitly measured attitudes continued to have a meaningful relationship with cultural representations even after controlling for participants' perceptions of how the culture evaluates topics. Implicitly measured attitudes thus appear to be tapping into the cultural representations that are relatively more hidden in language patterns and are not reducible to explicit, reportable cultural knowledge.

Implications for Perspectives on Implicit Social Cognition and the Persistence of Culture

Links between implicitly measured attitudes and cultural representations persisted over 200 years of books text and text from 29 non-English languages. Such evidence suggests that implicitly measured attitudes are affected by *chronically and widely* accessible cultural representations. Importantly, this idea that implicitly measured attitudes pick up chronically accessible content is unlike the analogy in the Bias of Crowds model that implicitly measured attitudes mainly pick up temporary culture as though it were a temporary "wave passing through a crowd" at a sports game (Payne et al., 2017). Rather, cultural representations in language appear to reflect an environment that is much more static and consistent—more like the sports arena

itself (similar ideas are evoked in the work by Payne et al., 2019).

To be clear, there is surely a role for temporary concept accessibility to explain temporary malleability in individual participants' implicitly measured attitudes (Kurdi & Charlesworth, 2023). However, when it comes to implicitly measured attitudes aggregated across large samples of participants, it appears that they better reflect the chronically accessible, persistent, and widespread concepts in culture. This interpretation also aligns with data showing that implicitly measured attitudes are slower to change over the long term (Charlesworth & Banaji, 2022a), because to change them would require uprooting long-enduring patterns and representations in culture and language (Charlesworth & Hatzenbuehler, 2024).

Implications for Understanding Explicitly Measured Attitudes

Explicitly measured attitudes were robustly distinct from cultural representations, with a meaningful absence of relationships not only in contemporary English but across all 200 years of text, and all 53 non-English languages. This is perhaps surprising given that implicitly and explicitly measured attitudes are often related to one another (Nosek, 2005, 2007). However, this implicit–explicit relationship is not perfect. The current data suggest that this imperfect overlap may arise because implicitly and explicitly measured attitudes are shaped from different sources. Whereas implicitly measured attitudes reflect hidden patterns of culture in language, explicitly measured attitudes may be more filtered through participants' personal values and ideologies and the norms about what they think they *should* express (Crandall et al., 2002; Devine, 1989).

Looking at example topics that show the strongest deviations helps reinforce this interpretation. Participants expressed an explicit preference for United States (over Japan), Science (over Astrology), and Intelligent (over Athletic people), all of which would be preferences that are in line with what "should be" expressed based on status and prestige (e.g., it is unpopular to like Astrology). Yet these topics showed the reverse associations in contemporary English and in implicitly measured attitudes (e.g., Astrology was more positive than Science on the IAT and in language). Although early perspectives (e.g., Rudman, 2004) proposed similar ideas that implicitly and explicitly measured attitudes come from unique sources, the current work provides some of the first quantitative large-scale data supporting such hypotheses about the differences of such attitudes.

Limitations

By necessity, the current project simplified the vast complexity of both attitudes and culture. First, we focused on a relatively small sample of 55 attitude topics that does not fully capture the variety of humans' attitudes. Although the relationship of implicitly measured attitudes and culture was robust across three topic domains (e.g., concrete objects, abstract concepts, and social groups), future work will nevertheless benefit from measuring attitudes and cultural representations across a broader sample space. This is especially important for historical and non-English analyses because attitude topics more relevant in the past (e.g., explorers, pioneers, or pirates) or to other cultures (e.g., racial groups such as Mulatto or Indigenous that matter more in other countries) were omitted because of the current focus on contemporary U.S. English data.

Second, and related, given the thousands of words to translate, we relied on ChatGPT3.5 to automate translation. However, the limitations of ChatGPT in generating false information (Alkaissi & Mcfarlane, 2023) could have introduced errors. Supplemental tests using hand-translated stimuli lists with fluent speakers suggested such errors were unlikely. Nevertheless, translations are generally limited because some concepts in English do not have direct analogues in non-English languages (and vice versa). Future research may therefore consider more culturally applicable translations and topics.

Third, all analyses were correlational. This was necessary for testing relationships between implicitly measured attitudes and language across long historical timespans (after all, we could not experiment on historical minds), and to scale-up analyses across languages and media types (e.g., internet, books, newspapers). Nevertheless, to improve causal interpretation, future work could add experimental approaches (e.g., Hauser & Schwarz, 2022), such as testing whether repeated exposure to texts with certain latent cultural representations (e.g., books reflecting more pro-Poor/anti-Rich representations) shift readers' implicitly but not explicitly measured attitudes.

Conclusion

Despite these limitations, the work delivered new insights into the nature of attitudes and their intertwining with cultural representations through language. Returning to the analogies offered in the opening citation from the work by Olson and Kenrick (2011), it seems clear that our culture, like a hidden environmental pollutant, "contaminates" our implicitly measured attitudes (and vice versa). Our more controlled and conscious explicitly measured attitudes, in contrast, appear to add a filter (perhaps personal values, ideologies, and norms of expression) such that pollutants are not directly reflected in explicit measures. Understanding this deep interweaving with culture is important for thinking about the possibility of, and optimal methods for, change. After all, although some of the attitude topics studied here are innocuous (e.g., Simple vs. Difficult, Hot vs. Cold), others carry social consequences (e.g., Straight vs. Gay, White vs. Asian) that warrant

change. The current work suggests that changing such consequential attitudes likely demands creative new strategies (Paluck et al., 2021) to simultaneously address not only the attitudes but their persistent and widespread links with cultural representations in language.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a *Social Sciences and Humanities Research Council* of Canada Postdoctoral Fellowship awarded to Tessa Charlesworth.

ORCID iD

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

Data Availability

All data and analyses are available at https://osf.io/9kg5h/. Analyses on contemporary data were preregistered at https://osf.io/d63pt

Supplemental Material

The supplemental material is available in the online version of the article.

Notes

1. We use the term implicitly measured attitudes to highlight that we are interpreting results in terms of the outcomes of measurement procedures (i.e., representations that are measured indirectly), and are more agnostic as to the construct itself (De Houwer et al., 2009).
2. For non-English analyses, we only perform the *WEAT D-short* analysis since (a) all metrics showed consistent correlations with one another and (b) we do not have data on valence ratings of traits in other languages, which would be required for MAC analyses.
3. Language families and geographic proximity are meaningfully related to one another, $r = .45$ [.21, .64], but are not perfectly overlapping (e.g., some physically close languages, such as Hungarian and Tatar, are from non-Indo European families). As such, we explore both proximity and language families as potential moderators.

References

Albarracín, D., Johnson, B. T., & Zanna, M. P. (2005). The handbook of attitudes. In D. Albarracin, B. T. Johnson, & Mark P. Zanna (Eds.), *The handbook of attitudes*. Lawrence Erlbaum. https://doi.org/10.4324/9781410612823

Alkaissi, H., & Mcfarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, *15*(2), Article e35179. https://doi.org/10.7759/cureus.35179

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 117–150). American Psychological Association. https://doi.org/10.1037/10394-007

Bhatia, S., & Walasek, L. (2023). Predicting implicit attitudes with natural language data. *Proceedings of the National Academy of Sciences*, *120*, Article e2220726120. https://doi.org/10.1073/pnas.2220726120

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Proceedings of the 30th Conference on Neural Information Processing Systems (pp. 4356–4364). New York, NY: Curran Associations Inc.

Bruner, E. M. (1956). Cultural transmission and cultural change. *Southwestern Journal of Anthropology*, *12*(2), 191–199. https://www.jstor.org/stable/3629114

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. (2022a). Patterns of implicit and explicit attitudes: IV. change and stability from 2007 to 2020. *Psychological Science*, *33*(9), 1347–1371. https://doi.org/10.1177/09567976221084257

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. 594–608). Guilford Publications.

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

Charlesworth, T. E. S., & Hatzenbuehler, M. L. (2024). Mechanisms Upholding the Persistence of Stigma across 100 Years of Historical Text. *Scientific Reports, 14*. https://doi.org/https://www.nature.com/articles/s41598-024-61044-z

Charlesworth, T. E. S., Sanjeev, N., Hatzenbuehler, M. L., & Banaji, M. R. (2023). Identifying and predicting stereotype change across 72 groups, four text sources, and historical time (1900-2015): Insights from word embeddings. *Journal of Personality and Social Psychology*, *125*(5), 969–990. https://doi.org/https://doi.org/10.1037/pspa0000354

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

Cunningham, W. A., Zelazo, P. D., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition*, *25*(5), 736–760. https://doi.org/10.1521/soco.2007.25.5.736

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347–368. https://doi.org/10.1037/A0014211

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5

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

Durkheim, E. (1974). *Sociology and philosophy*. The Free Press.

Fine, G. A. (1979). Small groups and culture creation: The idioculture of little league baseball teams. *American Sociological Review*, *44*(5), 733–745.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). *Package "rstanarm" type package title Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm

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

Greenhill, S. J., Currie, T. E., & Gray, R. D. (2009). Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society B: Biological Sciences*, *276*(1665), 2299–2306. https://doi.org/10.1098/RSPB.2008.1944

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1489–1501). Association for Computational Linguistics. https://doi.org/10.18653/v1/p16-1141

Hauser, D. J., & Schwarz, N. (2022). Implicit bias reflects the company that words keep. *Frontiers in Psychology*, 13, Article 871221. https://doi.org/10.3389/FPSYG.2022.871221/BIBTEX

Hussey, I., Hughes, S., Lai, C. K., Ebersole, C. R., Axt, J., & Nosek, B. A. (2012). *The Attitudes, Identities, and Individual Differences (AIID) study and dataset*. https://doi.org/10.17605/OSF.IO/PCJWF

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the implicit association test. *Journal of Personality and Social Psychology*, *81*(5), 774–788. https://doi.org/10.1037/0022-3514.81.5.774

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

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Kurdi, B., & Charlesworth, T. E. S. (2023). The 3D framework of implicit attitude change: Levels of analysis, sources of change, and timescales. *Trends in Psychological Science*, *27*(8), 745–758. https://doi.org/10.1016/j.tics.2023.05.009

Kurdi, B., Ratliff, K. A., & Cunningham, W. A. (2021). Can the implicit association test serve as a valid measure of automatic cognition? A response to Schimmack (2021). *Perspectives on Psychological Science*, *16*(2), 422–434. https://doi.org/10.1177/1745691620904080

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, *4*(10), 1021–1028. https://doi.org/10.1038/s41562-020-0918-6

Marsden, P. V., Smith, T. W., & Hout, M. (2020). Tracking US social change over a half-century: The general social survey at fifty. *Annual Review of Sociology*, *46*, 109–134. https://doi.org/10.1146/annurev-soc-121919-054838

Mohr, J., Bail, C. A., Frye, M., Lena, J. C., Lizardo, O., McDonnell, T. E., Mische, A., Tavory, I., & Wherry, F. F. V. (2019). *Measuring culture*. Columbia University Press. https://doi.org/10.7312/MOHR18028/HTML

Moscovici, S. (1994). Social representations and pragmatic communication. *Social Science Information*, *33*(2), 163–177. https://doi.org/10.1177/053901894033002002

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565–584. https://doi.org/10.1037/0096-3445.134.4.565

Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science*, *16*(2), 65–69. https://doi.org/10.1111/j.1467-8721.2007.00477.x

Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition & Emotion*, *22*(4), 553–594. https://doi.org/10.1080/02699930701438186

Olson, M. A., & Kendrick, R. V. (2011) Origins of Attitudes. In Attitudes and Attitude Change (Eds. William D. Crano & Radmila Prislin). (pp. 111–130). New York: Psychology Press.

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology*, *72*, 533–560. https://doi.org/10.1146/annurev-psych-071620-030619

Payne, B. K., Vuletich, H. A., & Brown-Iannuzzi, J. L. (2019). Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(24), 11693–11698. https://doi.org/10.1073/pnas.1818816116

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*(4), 233–248. https://doi.org/10.1080/1047840X.2017.1335568

Rudman, L. A. (2004). Sources of implicit attitudes. *Current Directions in Psychological Science*, *13*(2), 79–82. https://doi.org/10.1111/j.0963-7214.2004.00279.x

Schimmack, U. (2021). The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science*, *16*(2), 396–414. https://doi.org/10.1177/1745691619863798

van Bavel, J. J., Jenny Xiao, Y., & Cunningham, W. A. (2012). Evaluation is a dynamic process: Moving beyond dual system models. *Social and Personality Psychology Compass*, *6*(6), 438–454. https://doi.org/10.1111/j.1751-9004.2012.00438.x

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Author Biographies

Tessa E. S. Charlesworth is an Assistant Professor at the Kellogg School of Management at Northwestern University. She completed her PhD at Harvard University, and a Social Sciences and Humanities Research Council postdoctoral fellowship at the University of Toronto. Her research uses computational and big data methods to understand long-term change in social attitudes and stereotypes.

Kirsten Morehouse is a Psychology PhD student at Harvard University, funded by the National Science Foundation. She studies implicit stereotypes that conflict with ground-truth data (e.g., Human = White) and how implicit bias is propagated by generative AI.

Vaibhav Rouduri is a Software Engineer at Genpact and will soon be starting his Master's Degree in Computer Science at New York University. He completed his undergraduate degree in Computer Science from BITS Pilani University in India.

William Cunningham is a Professor of Psychology and Computer Science at the University of Toronto. He completed his PhD at Yale University. His research builds computational models of cognition to understand social behaviour.

Handling Editor: Adam Hahn