The relationship between implicit intergroup attitudes and beliefs

Benedek Kurdi[a,1], Thomas C. Mann[a], Tessa E. S. Charlesworth[a], and Mahzarin R. Banaji[a,1]

[a]Department of Psychology, Harvard University, Cambridge, MA 02138

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2018.

Contributed by Mahzarin R. Banaji, January 25, 2019 (sent for review November 28, 2018; reviewed by Alex Madva and Duane T. Wegener)

Intergroup attitudes (evaluations) are generalized valence attributions to social groups (e.g., white–bad/Asian–good), whereas intergroup beliefs (stereotypes) are specific trait attributions to social groups (e.g., white–dumb/Asian–smart). When explicit (self-report) measures are used, attitudes toward and beliefs about the same social group are often related to each other but can also be dissociated. The present work used three approaches (correlational, experimental, and archival) to conduct a systematic investigation of the relationship between implicit (indirectly revealed) intergroup attitudes and beliefs. In study 1 (n = 1,942), we found significant correlations and, in some cases, evidence for redundancy, between Implicit Association Tests (IATs) measuring attitudes toward and beliefs about the same social groups (mean r = 0.31, 95% confidence interval: [0.24; 0.39]). In study 2 (n = 383), manipulating attitudes via evaluative conditioning produced parallel changes in belief IATs, demonstrating that implicit attitudes can causally drive implicit beliefs when information about the specific semantic trait is absent. In study 3, we used word embeddings derived from a large corpus of online text to show that the relative distance of 22 social groups from positive vs. negative words (reflecting generalized attitudes) was highly correlated with their distance from warm vs. cold, and even competent vs. incompetent, words (reflecting specific beliefs). Overall, these studies provide convergent evidence for tight connections between implicit attitudes and beliefs, suggesting that the dissociations observed using explicit measures may arise uniquely from deliberate judgment processes.

attitudes | Implicit Association Test | implicit social cognition | stereotypes | word embeddings

The conceptual distinction between facts and preferences seems so obvious that even preschoolers recognize it without any difficulty: If Naomi thinks that germs are big and Jonah thinks that germs are small, most 5-y-olds will agree that they cannot both be right; however, if Mirabel's favorite color is blue and Isaac's favorite color is red, most 5-y-olds will agree that they can both be right (1). Nevertheless, especially in the context of social groups, the idea of a simple dichotomy between attitudes and beliefs has been replaced by an understanding that the two are both separate and deeply intertwined (2, 3). An intergroup attitude is usually defined as an evaluation of a group along a positive–negative continuum, whereas a belief (or stereotype) about a group is usually defined as consisting of specific semantic content that is not reducible to a positive–negative dimension. For instance, a preference for Asian Americans over another group on measures of liking, pleasantness, and warmth would be considered an attitude; attributions of traits like intelligence, diligence, and honesty to Asian Americans would be considered beliefs. As such, attitudes are conceptualized to be general, subjective, and valence-based, and beliefs are conceptualized to be specific, truth-evaluable, and semantically based.

However, such clear-cut separation of the human mind into attitudes or evaluative representations, on the one hand, and beliefs or semantic representations, on the other hand, may be rooted in phenomenology rather than in empirical evidence (4, 5). Pioneering work on the measurement of word meaning from the 1950s demonstrated that attitudes (valence) and beliefs (semantics) are inextricably connected (6): When dimension reduction techniques, such as factor analysis, are applied to the space of word meanings, the latent factor accounting for the majority of semantic variance is valence; that is, the evaluative component dominates word meaning. This analysis raises a fundamental question about the organization of social group representations: Can beliefs become dissociated from attitudes and, if so, under what conditions?

The Attitude–Belief Relationship: Evidence from Explicit Measures of Social Cognition

Although social group beliefs or stereotypes can vary from being relatively evaluatively neutral, such as "African Americans are taller than Asian Americans," to being strongly evaluative, such as "Asians are more deceitful than Europeans," the majority of social group beliefs are valenced. On average, participant-generated stereotypic traits of white, black, Hispanic, and Asian Americans (7) deviate by 1.68 points from the neutral midpoint of a nine-point valence scale, compared with an average deviation of 1.03 points for the entire English lexicon (8). Additionally, in the first known empirical study on social group stereotypes, numerous traits endorsed

Significance

Relying on evidence from explicit (self-report) measures, attitudes toward social groups ("I like Asians"), and beliefs about social groups ("I think Asians are smart") have been seen as conceptually and empirically distinct. Here, we use an experimental measure of response latency and an archival measure of textual distance to show that automatic attributions of competence to social groups (implicit beliefs) are both predicted by and causally related to automatic attributions of general positivity (implicit attitudes). Theoretically, these results suggest that attitude–belief dissociations may be a unique product of deliberate cognitive processes. Moreover, they raise the intriguing possibility that targeting implicit attitudes, whether in individual minds or in language, may result in concomitant shifts across a range of implicit beliefs.

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

Reviewers: A.M., California State Polytechnic University; and D.T.W., The Ohio State University.

The authors declare no conflict of interest.

Data deposition: The R code for all analyses, data files (including trial-level Implicit Association Test data), and materials are freely available from the Open Science Framework (https://osf.io/xyhgu/).

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

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1820240116/-/DCSupplemental.

Published online March 4, 2019.

by American participants were deeply evaluative: Germans were described as intelligent, Italians as lazy, African Americans as dirty, the Irish as pugnacious, and the English as honest (9). Furthermore, different social groups often anchor at opposite extremes of the valence spectrum. For instance, white Americans are stereotyped as relatively intelligent, wealthy, trustworthy, and friendly, whereas black Americans and Hispanic Americans are stereotyped as relatively unintelligent, poor, untrustworthy, and threatening (7).

As such, it may not be surprising that decades of research using explicit (self-report) measures of cognition have generated robust evidence for associations between attitudes and beliefs, including in the domain of social group representations (10–12). For instance, increased liking of African Americans has been shown to predict increased endorsement of positive traits, such as proud, and decreased endorsement of negative traits, such as lazy (11). Moreover, experimental studies have provided evidence that attitudes can even causally drive beliefs. Attitude shifts induced via evaluative conditioning have been demonstrated to lead to concomitant changes in group beliefs along the trait dimension of warmth (13). Given this evidence, social group stereotypes may be expected to be universally congruent with attitudes: Positive evaluations of a group should align with positive stereotypes, and negative evaluations of a group should align with negative stereotypes.

However, evidence for attitude–belief consistency is not without exceptions. When measured using self-report, attitudes can be self-contained and need not align with stereotypes. For instance, Asian Americans are subject to relatively negative explicit attitudes by white Americans but are positively stereotyped as intelligent (14). Conversely, the elderly can be subject to positive explicit attitudes but are negatively stereotyped as cognitively and physically slow (15). Moreover, explicit intergroup attitudes and explicit intergroup beliefs need not be correlated with each other (10, 11, 16). For instance, attitudes toward African Americans and endorsement of clearly valenced group stereotypes, such as superstitious, prejudiced, or oversensitive, have been found to be uncorrelated (11).

Indeed, self-reported group beliefs are often based on consensual cultural stereotypes, and thus need not depend on group membership or even the individual's specific attitudes. For instance, a white American respondent may endorse the stereotype that Asian Americans (an outgroup) possess a positive trait, such as intelligence (17), while white Americans (the ingroup) possess a negative trait, such as arrogance (7). Endorsement of such negative ingroup stereotypes and positive outgroup stereotypes may occur even though white Americans tend to exhibit a positive attitude toward their ingroup (18). Beyond these individual examples, tests of the stereotype content model also reveal consistent consensus among both American (17) and international (19) participants in self-reported social group stereotypes along the central dimensions of warmth and competence. As such, the stereotype content model provides compelling evidence that when explicit measures are used, beliefs about the specific traits of social groups can become dissociated from generalized group evaluations.

Finally, experimental studies have revealed that the acquisition of explicit stereotypes can be subject to validation processes; that is, controlled reasoning may be employed to ascertain whether a certain proposition logically entails another proposition (20–22), thereby constraining attitude–belief consistency. For instance, manipulating attitudes has been found to result in concurrent shifts in group stereotypes on warmth (a highly valenced trait that is itself often used as a measure of attitude) but not on competence (13). The reason for a lack of belief revision on the latter dimension is most likely that participants find it more reasonable to infer from the proposition "X is paired with pleasant images" the proposition "X is good" than the proposition "X is smart." The weak basis for drawing inferences about competence from the pairings can discourage participants from expressing revised beliefs about the intelligence trait.

The Attitude–Belief Relationship: Evidence from Implicit Measures of Social Cognition

Over the past decades, much research on intergroup cognition has been guided by the recognition that attitudes (social group evaluations) and beliefs (social group stereotypes) can be activated automatically upon encountering a stimulus (23–26). Such implicit attitudes and implicit beliefs can be measured unobtrusively using response latency tasks such as the Implicit Association Test (IAT) (27), as opposed to explicit attitudes and beliefs that are measured using self-report. In the present paper, we use the term implicit belief to refer to any automatically activated mental representation whose meaning cannot be reduced exclusively to valence. Use of the term implicit belief should not be taken as a sign of an a priori commitment to the idea that implicit beliefs are represented as propositions or that they are truth-evaluable (28). Similarly, our occasional use of the terms associate and association does not indicate an a priori commitment to associative theories (21, 22); rather, these terms are used as shorthand for automatically revealed conceptual compatibility, be it in the form of a proposition (e.g., "Asians are smart") or a mere association (e.g., Asian–smart). Surprisingly, beyond limited tests involving individual social groups and individual belief attributes (29–32), the relationship between implicit attitudes and implicit beliefs has never been systematically investigated (5).

As such, the studies addressing the implicit attitude–belief relationship reported here stand to provide novel insight into the basic structure of social group representations. Beyond obvious differences in methods of measurement, implicit cognition and explicit cognition have been posited to differ from each other in how directly they tap into evaluative mental content (21, 22). Specifically, implicit measures have been hypothesized to more directly index evaluative representations than their explicit counterparts (33–35). For instance, in the context of social groups, explicit measures may reflect certain kinds of propositional judgment, such as rejecting a negative evaluation of a group based on egalitarian ideals, self-presentational concerns, or knowledge of past suffering (36); implicit measures, on the other hand, may be less sensitive to such processes (20). Moreover, given that automatic responses often guide decisions about approaching or avoiding a stimulus, such responses may preferentially track evaluations along a positive–negative continuum (37) even if the measure nominally taps a specific belief (but ref. 38). If this is indeed the case, then implicit measures may reveal more consistent evidence in favor of attitude–belief associations than explicit measures.

In line with this idea, implicit and explicit attitudes have been shown to be more highly correlated with each other when participants were instructed to focus on their feelings rather than their thoughts (39–41). Moreover, at least under some conditions, implicit beliefs seem to be more closely related to explicit attitudes than they are to explicit beliefs (42). Finally, a recent investigation using a reinforcement learning perspective has found that whereas implicit attitudes track a history of personally experienced rewards and punishments, explicit attitudes also reflect additional information about the structure of the environment (43). However, it should be noted that most of these results have been obtained using implicit measures of attitude and, as such, may not readily generalize to implicit measures of beliefs. In addition, a competing theoretical perspective posits that the same set of propositional representations underlie responding on explicit and implicit measures (28). Under this position, there is no reason to expect any discrepancy between explicit and implicit measures in terms of the attitude–belief relationship.

Beyond their theoretical import, the current studies also bear on issues of (*i*) measurement, (*ii*) interventions designed to create change in implicit attitudes and beliefs, and (*iii*) the ecological validity of implicit measures. First, if the present studies were to demonstrate tight associations between implicit attitudes and implicit beliefs, such a result would have far-reaching implications for the interpretation of studies using implicit belief attributes highly discrepant in valence (e.g., smart vs. dumb). Second, implicit attitudes and beliefs have been shown to predict intergroup discrimination above and beyond their explicit counterparts (44). As such, designing interventions to produce change in implicit attitudes and beliefs has been an important endeavor in social cognition research. The present results can inform such interventions by shedding light on whether shifting attitudes and beliefs requires two separate sets of interventions (if implicit attitudes and beliefs are found to be empirically dissociable) or a unitary strategy (if implicit attitudes and beliefs are highly overlapping). Finally, to circumvent the methodological limitations inherent in the IAT, the present project also relied on measures of valence and trait attribution derived from a vast online corpus of spontaneously generated text data (45, 46). Convergence between results obtained using implicit measures administered to individual participants, on the one hand, and using word embeddings derived from a repository of public discourse, on the other hand, should increase confidence in the generalizability of the present findings beyond a single method. Moreover, such results from word embeddings would provide further evidence for a close connection between social group representations residing in individual minds and regularities in natural language (45).

Study 1

Study 1 tested the relationship between implicit attitudes and implicit beliefs using a correlational method involving several social group targets, belief attributes, and participant groups. Whereas evidence obtained with explicit (self-report) measures has demonstrated both associations (10–12, 47) and dissociations (10, 11, 16, 48) between attitudes and beliefs, the relationship between implicit attitudes and implicit beliefs has never been systematically investigated. As such, the present study provides an examination of the basic organization of automatically activated social group representations across multiple social group targets.

Design. Participants (final n = 1,942) completed two IATs (27) measuring implicit attitudes (generalized evaluations) and implicit beliefs (specific trait attributions) involving the same social group targets. The IAT measures the conceptual compatibility of two categories (e.g., Asian and white) and two attributes (e.g., smart and dumb) by comparing average response latencies across two speeded sorting tasks: a first sorting task in which one category and one attribute (e.g., Asian and smart) share a response key and the other category and other attribute (e.g., white and dumb) share the other response key, and a second sorting task in which the mapping of categories to attributes is reversed (e.g., Asian–dumb, white–smart). Further details on the method are provided in *SI Appendix*.

The specific target groups and traits used, along with participants' group membership, are displayed in Figs. 1 and 2. Attitudes and warmth stereotypes involve the same constituent elements (self + group X + warmth) and differ only in the direction of the relationship between the elements ("I feel warmly toward group X" vs. "group X feels warmly toward me"). As such, attitudes and warmth stereotypes may not be easily distinguishable from each other using traditional implicit measures, such as the IAT. To allow for attitude–belief dissociations to emerge, the stereotypes used in study 1 were all sampled from the domain of competence (17), including smart/dumb (studies 1A and 1B), book-smart/street-smart (study 1C), and mental/physical (studies 1D and 1E).

To ensure that the results are not a function of any specific belief, study 1 examined a variety of traits that differed in their evaluative content. The traits used in studies 1A and 1B were evaluatively discrepant, whereas those used in study 1C were evaluatively similar, and those used in studies 1D and 1E were seemingly evaluatively equivalent. Conducting an initial study with belief attributes that are inherently valenced, one highly positive (smart) and the other highly negative (dumb), is of interest for three reasons. First, social group stereotypes tend to be strongly valenced (7, 9, 17). As such, to provide an accurate reflection of the domain, this initial test was performed involving a pair of highly valenced belief attributes. Second, explicit measures of intergroup cognition have provided evidence for attitude–belief dissociations even when the belief is highly valenced, and hence should share an evaluative component with the measure of attitude (10, 11, 14–17, 48). Similarly, if implicit attitudes and beliefs were found to be unrelated even when the beliefs are highly valenced, this pattern of results

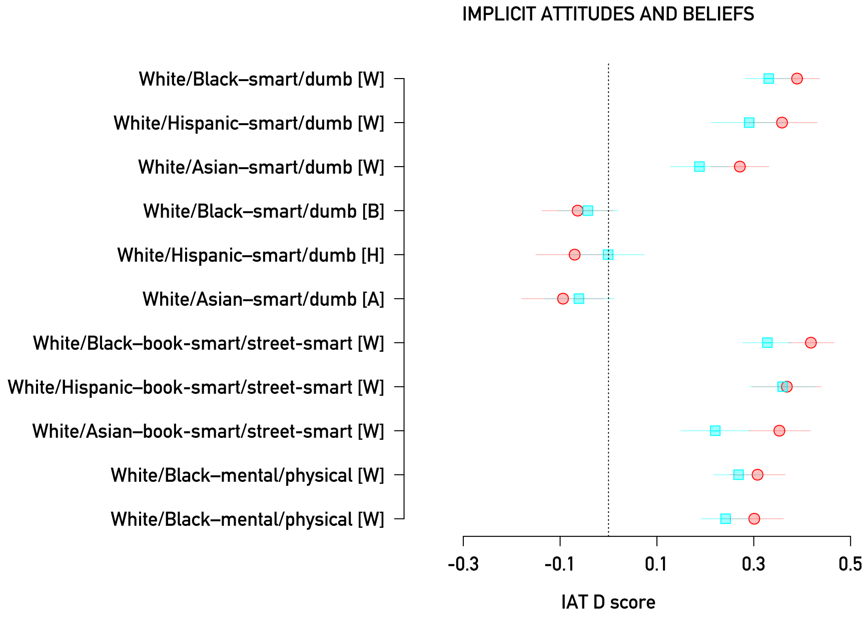

Fig. 1. Mean implicit attitudes and implicit beliefs obtained in study 1 (n = 1,942). The x axis shows IAT D scores indexing the strength of implicit attitudes and beliefs such that stronger prowhite attitudes and stronger association of the positive belief attribute with white are reflected by positive D scores. The y axis provides information about the target groups (e.g., white/black) and belief attributes (e.g., smart/dumb), with participant race in brackets (A, Asian; B, black; H, Hispanic; W, white). Red circles show attitudes, and blue squares show beliefs. Error bars show 95% CIs.

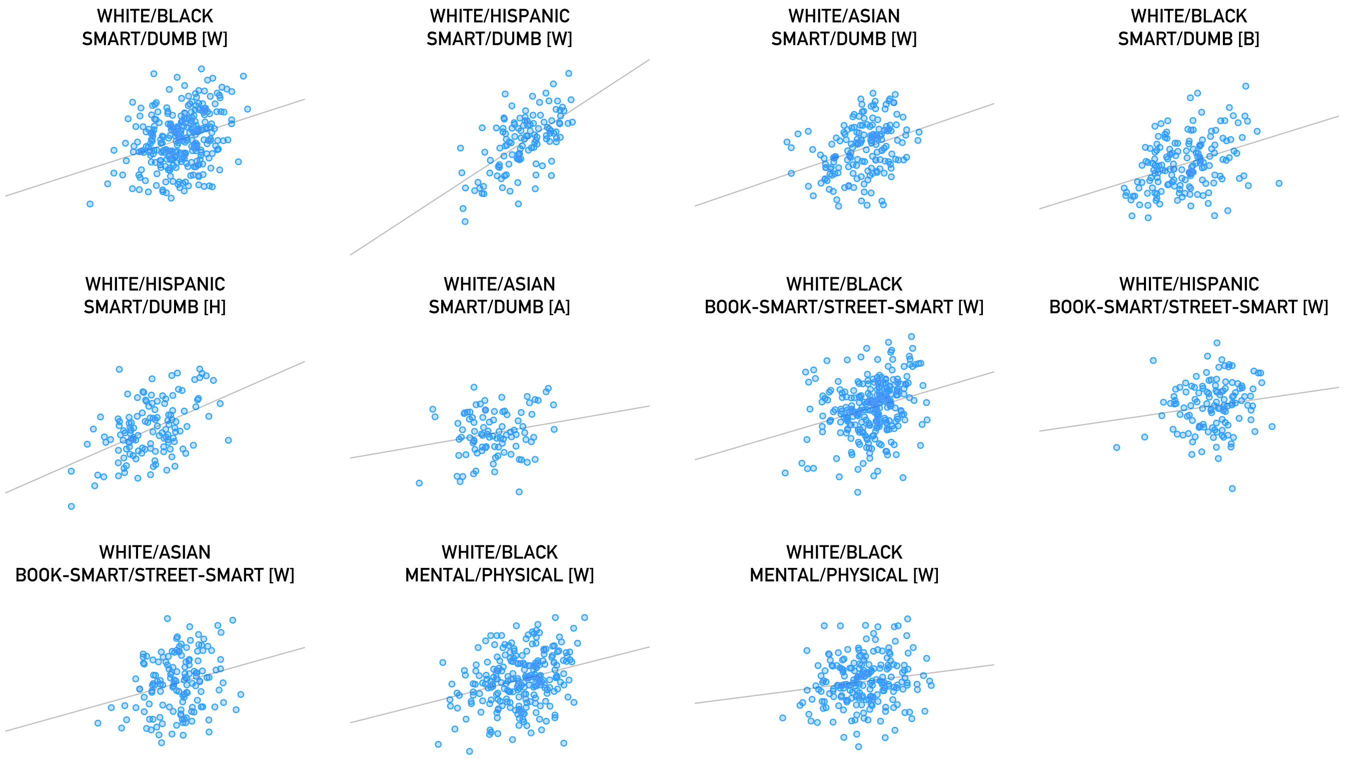

Fig. 2. Implicit attitude–belief correlations obtained in study 1 (n = 1,942). The x axis shows IAT D scores indexing the strength of implicit attitudes, and the y axis shows IAT D scores indexing the strength of implicit beliefs. Target groups and belief attributes (e.g., white/black–smart/dumb) are indicated above each plot, with participant race in brackets (A, Asian; B, black; H, Hispanic; W, white).

would provide especially strong evidence in favor of a dissociation. Finally, much previous research has tacitly assumed that IATs using highly valenced traits (e.g., smart vs. dumb) allow for inferences about implicit beliefs with specific semantic content beyond valence; however, this assumption has rarely been directly investigated (but refs. 49, 50).

However, extreme valence differences between trait dimensions (e.g., smart vs. dumb) may artificially inflate attitude–belief associations even if attitudes and beliefs are generally conceptually distinct from each other. Therefore, it is important to test whether attitude–belief consistency would also emerge if the belief attributes were more evaluatively equivalent. Accordingly, in study 1C, we used two relatively positive belief attributes (book-smart vs. street-smart). Finally, studies 1D and 1E were designed to provide an even more conservative test of attitude–belief consistency: The belief IAT in these studies used the traits mental vs. physical, and this implicit belief has been found to be uncorrelated with implicit attitudes in previous work (29). In study 1D, the original stimuli used by Amodio and Devine (29) were retained, whereas study 1E relied on a novel set of stimuli specifically created to minimize the attitude–belief correlation.

As a further test of the attitude–belief relationship, studies 1A and 1B included measures of the same beliefs (smart vs. dumb) and same targets (white vs. black, white vs. Hispanic, and white vs. Asian) but sampled participants from different social groups. Specifically, only white Americans were recruited for study 1A, and only black, Hispanic, and Asian Americans were recruited for study 1B. Previous theoretical (51) and empirical (52) work suggests that group membership is a major determinant of group evaluations, with high-status groups exhibiting ingroup preference and lower status groups exhibiting neutrality on implicit measures of attitude. If implicit measures of belief, like implicit measures of attitude, are largely reflective of generalized evaluative representations, they should reveal a similar pattern of ingroup-favoring beliefs among white participants and neutral beliefs among lower status participants. On the other hand, if implicit measures of belief, like explicit measures of belief, are responsive to consensual cultural stereotypes, a pattern of attitude–belief dissociation should be observed. For instance, relatively lower status groups may reveal neutral attitudes but outgroup-favoring beliefs (e.g., "white Americans are smarter than black Americans").

Results.

Group-level attitude–belief relationship. The results of study 1 provide robust evidence for group-level implicit attitude–belief consistency, as revealed by the alignment in the mean levels of attitudes and beliefs (Fig. 1). In study 1A, white participants exhibited attitudinal ingroup preference relative to all three comparison groups tested (black Americans, Hispanic Americans, and Asian Americans), thus replicating previous results (52). Moreover, in study 1B, participants sampled from these relatively lower status groups were, on average, found to be neutral on a measure of implicit attitudes, also replicating previous results (52).

When it comes to implicit beliefs, white participants in study 1A associated their ingroup relatively more with the positive trait smart compared with the negative trait dumb irrespective of the comparison group tested. This finding is unsurprising for the white/black and white/Hispanic contrasts considering that pervasive cultural stereotypes attribute more intelligence to white Americans than to black and Hispanic Americans (7, 17). However, the finding is unexpected for the white/Asian contrast, given that Asian Americans are consistently stereotyped as smart, and more so than white Americans (7, 14, 17). As such, this result provides a first indication that implicit beliefs among white Americans may be relatively insensitive to the semantic content of cultural stereotypes when such stereotypes favor a relatively lower status outgroup.

In study 1B, the same belief (smart vs. dumb) was tested among the members of lower status groups (black Americans, Hispanic Americans, and Asian Americans). Among these participants, attitude and belief means again aligned with each other, thus providing additional evidence for group-level attitude–belief consistency. Remarkably, such consistency emerged even though it contradicted established cultural stereotypes. Based on cultural stereotypes (7, 17), black and Hispanic participants could have been expected to associate the positive trait (smart) with the white outgroup and the negative trait (dumb) with the ingroup. By contrast, Asian participants may have been expected to associate the trait smart with their ingroup and the trait dumb with the outgroup (7, 14, 17). Instead, implicit beliefs were neutral in all three groups, tracking neutral implicit attitudes.

Study 1C tested a seemingly evaluatively neutral belief (book-smart vs. street-smart) among white American participants. Similar to study 1A, participants consistently associated book-smart with the ingroup and street-smart with the outgroup irrespective of the comparison group. This result is, again, unsurprising when white Americans are compared with black or Hispanic Americans, who are culturally stereotyped as street-smart rather than book-smart, especially relative to white Americans (7). However, as above, the same result is surprising when Asian Americans serve as the contrast category, given that the cultural stereotype of this group prominently features both academic excellence and a lack of street smarts (7, 14, 17). This finding suggests that, at least in the present context, book-smart may be a more genuinely positive trait than street-smart and, as such, may come to be associated with the high-status ingroup (a possibility that we explore in more detail below). [We do not wish to argue that book-smart is a universally more positive trait than street-smart. In fact, recent work has provided evidence that even traits with clear moral implications, such as merciful, honest, and selfish, are subject to contextually dependent implicit evaluations (53). As such, we believe that it is an intriguing empirical question whether a trait that becomes contextually positive given some current goal, also becomes temporarily more closely associated with the (high-status) ingroup.]

Finally, in studies 1D and 1E, white participants exhibited both attitudinal ingroup preference and a significant association of their ingroup with the trait mental and the black outgroup with the trait physical. This result is in line with previous work (29); however, given a lack of obvious valence difference between the two traits, its interpretation is ambiguous without considering the attitude–belief correlation (discussed below).

Individual-level attitude–belief relationship. In line with the results of the group-level tests reported above, we also found robust evidence for individual-level attitude–belief consistency, as revealed by correlations between attitude and belief IAT scores (Fig. 2). When the traits used on the belief IAT strongly differed in positivity (smart vs. dumb; studies 1A and 1B), moderate to large attitude–belief correlations were obtained. Remarkably, the attitude–belief correlation remained significantly positive even when both traits were seemingly positive (study 1C; book-smart vs. street-smart): White participants with higher levels of ingroup preference were more likely to associate book-smart with the ingroup. This suggests that, at least when evaluated automatically, book-smart may be a more genuinely positive trait than street-smart. (In line with this idea, the results of a supplementary study measuring implicit attitudes toward the traits book-smart vs. street-smart revealed robust implicit preference for the former over the latter [$t(406) = 12.16$, $P < 0.0001$, Bayes factor in favor of the alternative hypothesis (BF_{10}) = 8.81×10^{25}, Cohen's $d = 0.60$; *SI Appendix*].) Finally, the positive correlation emerged even when the traits were selected because they had revealed attitude–belief dissociations in prior research (29) or had even been normed to be evaluatively neutral using explicit measures of evaluation (mental vs. physical; studies 1D and 1E). Similar to study 1C, these results suggest that despite explicit neutrality, automatic evaluations of these traits may not be equally positive. (In line with this idea, the results of a supplementary study measuring implicit attitudes toward the traits mental vs. physical revealed robust implicit preference for the former over the latter [$t(372) = 14.22$, $P < 0.0001$, $\mathrm{BF}_{10} = 4.10 \times 10^{33}$, Cohen's $d = 0.74$; *SI Appendix*].) Taken together, these results provide reliable evidence for individual-level attitude–belief consistency across a diverse set of belief attributes and among members of both relatively higher status and lower status groups. Notably, such consistency emerged even when trait adjectives did not show blatant evaluative differences (study 1C) or had even been normed to be evaluatively equal (studies 1D and 1E).

However, one limitation of zero-order correlations as a measure of association is that they do not take into account measurement error in observed scores, which places an upper bound on the highest possible correlation attainable between attitude and belief IATs (*SI Appendix*). As such, zero-order correlations cannot speak to the issue of whether true scores in attitudes and beliefs are merely related to each other or are fully redundant. To investigate this question, we tested whether residual true variance in belief IAT scores remained significantly different from zero after partialing out (*i*) measurement error and (*ii*) true variance accounted for by attitude IATs (details of the procedure are provided in *SI Appendix*).

Using this approach, some implicit beliefs in studies 1A and 1B, which tested the most strongly valenced trait dimensions of any of the studies, were found to be redundant with implicit attitudes. Specifically, in both studies, residual true variance in implicit intelligence beliefs (smart vs. dumb) was not significantly different from zero for the white/Hispanic and white/Asian contrasts. On the other hand, residual true variance in belief IATs remained significantly different from zero for the white/black contrast, indicating that automatic intelligence attributions to these targets were related to, but not fully explained by, implicit attitudes.

Studies 1C–1E produced a different pattern of results. Unlike in studies 1A and 1B, no evidence for redundancy between implicit attitudes and implicit beliefs was obtained: A significant portion of variance in all belief IATs remained after accounting for measurement error and the effects of attitude IATs; that is, even though intergroup attitudes and intergroup beliefs were related to each other via shared valence, the former did not account for all true variance in the latter. Variance shared between attitudes and beliefs suggests that the traits used in studies 1C–1E (book-smart vs. street-smart and mental vs. physical) were not fully evaluatively equivalent. At the same time, the lack of complete redundancy between attitudes and beliefs suggests that these traits were nevertheless more evaluatively equivalent than the traits used in studies 1A and 1B (smart vs. dumb).

Study 2

Study 1 has provided robust evidence for attitude–belief consistency in implicit intergroup cognition at both the group level and the individual level. However, given its correlational design, study 1 cannot inform about the direction of a potential causal relationship between implicit attitudes and beliefs. It has already been demonstrated that, as expected, pairing group members with valenced belief attributes, such as strong vs. weak, shifts group evaluations on implicit measures of attitude (54, 55). Is learning in the opposite direction also possible? In other words, can repeatedly pairing members of a group with positive stimuli indirectly produce an association of that group with positively evaluated traits, such as American or mental? Conversely, can repeatedly pairing members of a group with negative stimuli indirectly produce an association with traits such as foreign or physical?

Study 2 investigated this question by probing whether implicit beliefs, such as automatic attributions of the traits American vs. foreign (study 2A) or mental vs. physical (study 2B), can shift in the face of a purely evaluative learning intervention, such as evaluative conditioning (56). A recent study conducted using explicit measures has provided mixed evidence on this issue:

Explicit beliefs about warmth, but not about competence, changed significantly as a result of evaluative conditioning (13).

Changes in implicit beliefs of nationality (American vs. foreign) and competence (mental vs. physical) via evaluative conditioning would offer particularly compelling evidence for the crucial role of implicit social group attitudes in driving implicit social group beliefs. First, beliefs about nationality are objectively verifiable, and validation processes should therefore lead to the inference that pairing a group with positive images is not diagnostic of that group's nationality (20–22). Second, automatic attributions of the traits mental vs. physical have traditionally been treated as the prime example for dissociations between implicit attitudes and beliefs (29), and therefore provide a notable test for examining the potential causal relationship between the two.

Design. The study consisted of a learning phase and a test phase. In a between-participant design, participants were randomly assigned to either an experimental condition or a control condition. In the experimental condition, members of one novel group (Niffians) were paired with positive images and members of another novel group (Laapians) were paired with negative images (57) semantically unrelated to nationality or intelligence. In the control condition, participants were exposed to the same number of stimulus pairings but group members were never paired with valenced images, thus providing a baseline measure of implicit attitudes and beliefs. In the test phase, implicit attitudes and implicit beliefs (American vs. foreign, study 2A; mental vs. physical, study 2B) were measured using an IAT (27).

Results. Attitude IAT scores significantly differed from each other across the control and experimental conditions [$t(156.01) = 5.70$, $P < 0.0001$, $\mathrm{BF}_{10} = 1.09 \times 10^6$, Cohen's $d = 0.78$ (study 2A) and $t(117.46) = 4.47$, $P < 0.0001$, $\mathrm{BF}_{10} = 1.59 \times 10^3$, Cohen's $d = 0.73$ (study 2B)], replicating well-established evaluative conditioning effects involving the same stimulus materials (57). Crucially, parallel changes in belief IAT scores were also observed. When Niffians were paired with positive images and Laapians with negative images that were semantically unrelated to the traits in question, Niffians became relatively more strongly associated with the trait American and Laapians with the trait foreign [$t(164.13) = 3.75$, $P = 0.0002$, $\mathrm{BF}_{10} = 1.44 \times 10^2$, Cohen's $d = 0.51$ (study 2A)] and Niffians became relatively more strongly associated with the attribute mental and Laapians with the trait physical [$t(129.61) = 3.35$, $P = 0.0011$, $\mathrm{BF}_{10} = 23.84$, Cohen's $d = 0.55$ (study 2B)].

These results suggest that purely evaluative information can drive implicit social group beliefs in the absence of specific semantic information about the dimension of interest. For the American/foreign contrast, this finding may be seen as particularly surprising because, unlike some cultural stereotypes involving trait attributions to social groups [e.g., Asian/cold (17)], beliefs about nationality are objectively verifiable. As such, processes of propositional validation (20–22) should have led to the conclusion that pairings of individuals with valenced images are not, in any way, diagnostic about whether the individuals are American or foreign. The fact that relative attributions of American vs. foreign shifted without relevant semantic information robustly demonstrates the causal power of evaluative learning in producing changes in implicit beliefs.

Significant changes in beliefs as a result of purely evaluative learning may also be seen as unexpected for implicit beliefs about the mental/physical dimension, because this contrast has been repeatedly used as a proof of concept for dissociations between implicit attitudes and implicit beliefs. Here, we have demonstrated not only a correlation of automatic attributions of mental vs. physical with implicit attitudes (studies 1D and 1E) but also a causal relationship: Pairings of group members with valenced images semantically unrelated to the relevant traits produced shifts in implicit mental/physical beliefs (study 2B).

Comparison of Implicit and Explicit Measures Across Studies 1–2

To compare the central tendency in implicit vs. explicit attitude–belief correlations across studies 1–2, two meta-analytic averages were computed (58). These measures revealed a high degree of similarity between implicit and explicit measures: The weighted mean attitude–belief correlation for implicit measures was $r = 0.36$, 95% confidence interval (CI): [0.28; 0.43], $P < 0.0001$, compared with $r = 0.32$, 95% CI: [0.23; 0.41], $P < 0.0001$, for explicit measures. For explicit measures, this result is in line with decades of theorizing and empirical results on individual-level attitude–belief consistency in the intergroup domain (10–12, 47). However, in a deviation from results obtained with implicit measures, explicit attitudes and beliefs were not uniformly consistent with each other at the individual level: Explicit beliefs about intelligence in the context of the white/Asian contrast were found to be uncorrelated with general group evaluations among both white ($\mathrm{BF}_{01} = 2.59$) and Asian ($\mathrm{BF}_{01} = 5.47$) participants. This finding raises the intriguing possibility that cultural stereotypes favoring lower status groups may be less likely to drive explicit intergroup attitudes than cultural stereotypes favoring higher status groups.

Explicit measures also diverged from implicit measures in terms of the effects of group membership. On implicit measures, white participants exhibited consistent ingroup preference and lower status participants exhibited consistent neutrality. However, when explicit measures were used, white American participants expressed significant association of the Asian outgroup with the traits smart (study 1A) and book-smart (study 1C) but revealed no significant deviation from neutrality in attitudinal preference. Crucially, Asian Americans expressed the same explicit belief as white Americans but differed from them on a measure of attitude (study 1B): They exhibited an explicit belief linking their ingroup to intelligence, along with attitudinal ingroup preference. Overall, these data suggest that group membership is a major determinant of implicit beliefs via its effect on group attitudes. By contrast, on explicit measures, the effects of group membership may be overridden by additional factors, such as shared cultural knowledge.

Finally, unlike with implicit measures, where evidence for change in group beliefs was unequivocal following a purely evaluative learning intervention, the results obtained using explicit measures were mixed. Explicit beliefs about the novel targets being American vs. foreign did not significantly differ across the control vs. experimental conditions [$t(109.62) = 1.92$, $P = 0.057$, $\mathrm{BF}_{10} = 1.47$, Cohen's $d = 0.31$ (study 2A)]. However, compared with the control condition, participants in the experimental condition showed stronger endorsement of the novel targets being relatively more intelligent rather than athletic [$t(83.65) = 3.41$, $P = 0.001$, $\mathrm{BF}_{10} = 1.16 \times 10^2$, Cohen's $d = 0.63$ (study 2B)]. Taken together, these results demonstrate selective operation of propositional validation processes in explicit social cognition as opposed to implicit social cognition, where no evidence of such processes was obtained. In line with the same idea, when novel groups were used as targets in studies 2A and 2B, attitude–belief correlations were higher using implicit measures rather than explicit measures, suggesting that on the latter, participants were more likely to reject purely evaluative information as a valid basis for responding on measures of belief.

Study 3

Studies 1–2 have revealed robust evidence of attitude–belief consistency when implicit measures were used to probe these constructs. However, it may be argued that such consistency may, at least in part, have emerged due to method-specific variance shared by the different IATs administered to participants (59) or other methodological features of the IAT.

As such, in study 3, we investigated the relationship between attitudes and beliefs using a method that shares no method variance with the IAT: word embeddings derived from a corpus of over 600 billion tokens (60). Specifically, we used relative textual distances to calculate the location of 22 social groups in a semantic space defined by warmth (i.e., friendliness), competence (i.e., intelligence), and valence (i.e., general positivity). Crucially, prior work relying on explicit measures has found that warmth and competence are orthogonal to each other (17). For instance, the elderly are commonly reported to be warm but incompetent, whereas professionals are reported to be cold but competent. Moreover, ratings on warmth and competence have been shown to be independent of participants' group membership and their generalized social group evaluations (17). By contrast, the present studies 1 and 2 have demonstrated tight connections between generalized valence attributions (attitudes) and attributions of specific traits to social groups (beliefs) when implicit measures are used.

In the face of this divergence between the current work and previous findings, study 3 can provide independent and ecologically valid evidence on the attitude–belief relationship and offer potential indications regarding the origin of social group representations revealed by explicit (self-reported) and implicit (automatic) measures. Although the textual data used to derive the word embeddings differ in many ways from responses on the IAT, both methods share a crucial similarity: In both cases, inferences about group-based attitudes and beliefs are made indirectly; that is, unlike with traditional explicit measures, the data do not emerge from deliberate judgments about the groups in question. On the other hand, the linguistic data underlying word embeddings are more similar to traditional explicit measures than implicit measures in that they were produced in a relatively controlled manner and may, at least in part, be subject to self-presentational concerns. As such, competing predictions can be derived: Data produced using word embeddings may more closely resemble the results using implicit measures obtained here, or results using explicit measures obtained in earlier work (17, 19). Notably, if the attitude–belief association found on the IAT were to replicate using this method, this finding would suggest that the attitude–belief dissociations revealed by explicit measures depend uniquely on deliberate cognitive processes that do not operate under time pressure (e.g., on the IAT) or under less tightly controlled conditions outside the laboratory (e.g., producing text online).

Design. Word embeddings use information about co-occurrences of words within textual data to compress complex word meaning into a space with limited dimensionality (e.g., using vectors of length 300) (60, 61). One major advantage of this method is that, unlike the text from which they are derived, word embeddings can be subject to mathematical operations. In particular, the cosine of the angle between two vectors can be used as a measure of semantic similarity: Vectors with similar orientations in semantic space can be interpreted as similar in meaning. Inspired by recent work that has relied on word embeddings to investigate social psychological phenomena (45, 46), we used word embeddings here to probe the relationship between three fundamental dimensions of social cognition: warmth (friendly vs. unfriendly) (17), competence (smart vs. dumb) (17), and valence (good vs. bad). Specifically, we calculated standardized effect sizes expressing the relative distance of 22 social group labels from (*i*) warm vs. cold words, (*ii*) competent vs. incompetent words, and (*iii*) positive vs. negative words in a 300D space derived from the Common Crawl corpus of over 600 billion tokens of online text using the fastText algorithm (60). (A list of all stimuli and details of how cosine distances and effect sizes were calculated are provided in *SI Appendix*.)

Results. The data obtained from an analysis of textual distances resembled the data obtained using implicit measures in studies 1 and 2: Unlike in investigations using explicit measures (17), the relative distance of social groups from valence in semantic space positively predicted their distance from both warmth and competence (Fig. 3). In other words, warmth and competence were significantly correlated with each other ($r = 0.74$, 95% CI: [0.47; 0.88], $P < 0.0001$, $\mathrm{BF}_{10} = 4.16 \times 10^2$). Similarly, significant correlations were found between valence and warmth ($r = 0.81$, 95% CI: [0.60; 0.92], $P < 0.0001$, $\mathrm{BF}_{10} = 5.53 \times 10^3$) and between valence and competence ($r = 0.77$, 95% CI: [0.52; 0.90], $P < 0.0001$, $\mathrm{BF}_{10} = 1.05 \times 10^3$). [To demonstrate discriminant validity, we also calculated correlations between arousal (high vs. low intensity) and warmth ($r = -0.17$, 95% CI: [−0.55; 0.27], $P = 0.451$, $\mathrm{BF}_{01} = 2.90$) and between arousal and competence ($r = 0.10$, 95% CI: [−0.34; 0.50], $P = 0.674$, $\mathrm{BF}_{01} = 3.48$), and found evidence for the absence of a relationship. Such lack of correlation was not due to insufficient reliability of the arousal measure (split-half correlation: $r = 0.80$). Details are provided in *SI Appendix*.] In *SI Appendix*, we also report additional analyses confirming the generality of these findings using the same algorithm but a different corpus and the same corpus but a different algorithm.

Discussion

The present project constitutes a systematic investigation of the relationship between implicit attitudes (automatic attributions of

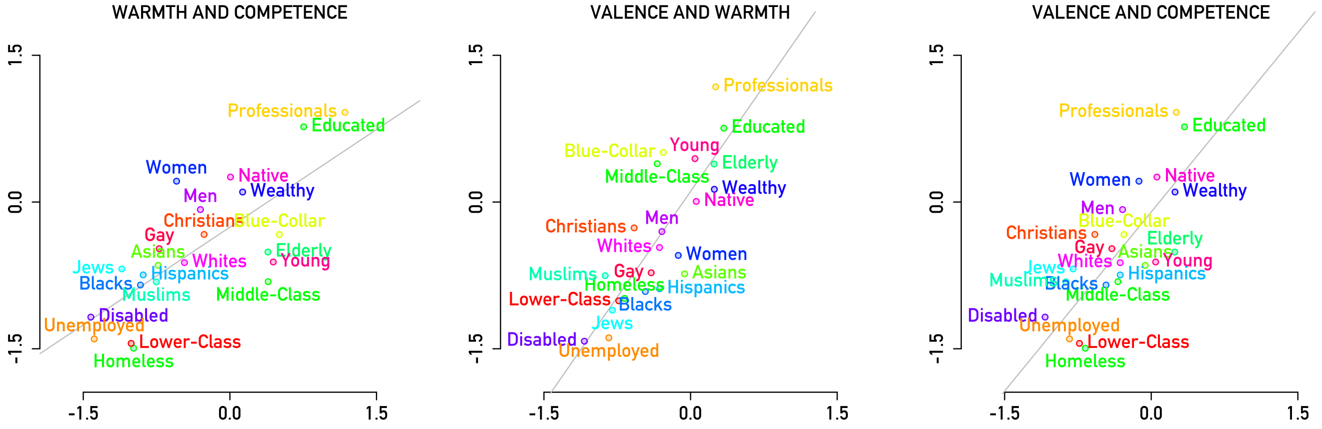

Fig. 3. Relationship among basic dimensions of social cognition obtained from word embeddings derived from the Common Crawl online corpus using the fastText algorithm (study 3). The *x* axis shows warmth, and the *y* axis shows competence (*Left*); the *x* axis shows valence, and the *y* axis shows warmth (*Center*); and the *x* axis shows valence, and the *y* axis shows competence (*Right*).

positive or negative valence to social groups) and implicit beliefs (automatic attributions of specific traits, such as smart, physical, or American, to social groups), extending across multiple social groups and multiple semantic dimensions. Across three studies, each relying on a different approach (correlational, experimental, or archival), robust evidence was found for consistency between implicit attitudes and implicit beliefs. Such attitude–belief consistency emerged even in cases where (*i*) parallel explicit attitudes and explicit beliefs revealed lack of consistency, (*ii*) it contradicted well-known cultural stereotypes (e.g., Asian American–smart), and (*iii*) propositional processes of validation should have constrained attitudinal valence from affecting measures of specific beliefs (e.g., beliefs about nationality).

Past work using explicit (self-report) measures to index social group representations has provided ample evidence for consistency between attitudes toward a social group and beliefs about the same social group (10–12). However, such measures have also repeatedly demonstrated dissociations. Specifically, positive attitudes toward a group need not imply positive evaluation of the same group along all trait dimensions (10, 11, 16). Moreover, participants often report positive explicit stereotypes of outgroups and negative explicit stereotypes of ingroups, resulting in consensual stereotypes spanning group boundaries (7, 9, 17). Finally, in studies of belief acquisition, propositional processes of validation (20) have been shown to operate on explicit beliefs. For instance, learning that a group is warm does not entail that the group is also competent, and explicit measures of group belief tend to reveal reasoning in line with such logical rules (13). The present work, although not designed with a primary focus on explicit measures of cognition, provides convergent evidence for each of these ideas.

By contrast, in the present project, implicit attitudes and implicit beliefs were found to be invariably consistent with each other, revealing (*i*) similar group means (including in tests involving participants from lower status and higher status racial groups); (*ii*) significant attitude–belief correlations; (*iii*) significant portions of variance, and sometimes all true variance, in beliefs accounted for by attitudes; and (*iv*) significant shifts in implicit beliefs as a result of purely evaluative learning. Moreover, despite cultural stereotypes to the contrary (7, 14, 17), white participants automatically attributed the traits smart and book-smart more strongly to their ingroup than to the Asian outgroup, whereas results among Asian participants revealed neutrality, with results in both groups tracking attitudes. Furthermore, implicit attitudes and implicit beliefs were found to be robustly correlated with each other, even in cases where explicit measures revealed dissociations (e.g., white/Asian–smart/dumb). Finally, and partly in contrast to explicit results, implicit attribution of positive traits, such as American and mental, to novel targets increased, whereas implicit attribution of negative traits, such as foreign and physical, decreased simply as a result of pairing group members with positive images semantically unrelated to the traits in question.

Accordingly, the present studies suggest that when it comes to implicit social cognition, evaluative equivalence between two traits is difficult to achieve, even for traits that seem, at first glance, to be equally positive. Mental and physical are not only semantically but also evaluatively opposed attributes: At least within the present context, mind is more positive than body, and it comes to be attributed to the ingroup to the extent that the ingroup is evaluated positively. This is also the case for book-smart and street-smart: Book-smart is more positive than street-smart, and hence more strongly attributed to the ingroup, even when such attribution contradicts the cultural stereotype. Related to this idea, the present project provides evidence that seemingly positive stereotypes associated with lower status groups, such as street-smart or athletic, are not as genuinely positive as stereotypes associated with higher status groups (62).

Taken together, these findings clearly support the core idea of theories of implicit social cognition that suggest that implicit attitudes and implicit beliefs are inextricably linked to each other due to the evaluative content that they share (5). At a more general level, the present work corroborates empirical work and theorizing on the major role of evaluation in shaping word meaning (6), automatic stimulus processing (33–35), and putatively purely cognitive higher level reasoning (4). Furthermore, the present results seem to be generally in line with dual-process theories of social cognition (20–22) in that they demonstrate a dissociation between explicit and implicit social cognition, with the former characterized by various constraints on attitude–belief consistency and the latter characterized by a lack of such constraints. Perhaps most relevant to the present project, it has been suggested that the ability to establish whether one proposition (e.g., "Niffians were paired with pleasant images") logically entails another proposition (e.g., "Niffians are American") is a unique feature of controlled processes of reasoning revealed by explicit (self-report) measures but not of automatic processes revealed by implicit (indirect) measures (20). The current results are in line with this position. [The present results may also be compatible with single-process theories, such as the propositional perspective on implicit evaluation (28). Specifically, the propositional model posits that the ability to encode relational qualifiers is the defining difference between associations and propositions, and the present work does not speak to this distinction. Under the propositional perspective, changes in implicit beliefs as a result of purely evaluative learning could be characterized as the output of "quick and dirty" propositional processes. Single-process perspectives positing that explicit measurement allows more time for a response to draw upon a wider set of information in memory, such as considerations of logical validity or nonevaluative information, are also consistent with the present findings.]

In addition to the theoretical implications discussed above, the present results are germane to the design of future studies using the IAT (27) with the goal of measuring implicit stereotypes. For instance, a recent study sought to investigate associations between native vs. foreign accent along multiple trait dimensions (trustworthy–untrustworthy, competent–incompetent, and social–unsocial) (63). However, the stimuli selected for the trustworthy, competent, and social attributes consisted exclusively of positive adjectives (e.g., "sincere," "capable," "warm"), and the stimuli selected for the untrustworthy, incompetent, and unsocial attributes consisted exclusively of negative adjectives (e.g., "deceitful," "incapable," "cold"). In addition to this example, numerous other IATs designed to assess implicit beliefs have used highly evaluatively discrepant attribute categories and stimuli (64–66). The findings of the present project, and especially studies 1A and 1B (using smart/dumb as the target attributes), caution against interpreting results obtained with such IATs as reflecting contentful implicit beliefs rather than merely differential attribution of positive vs. negative valence to the target groups. Further methodological implications are discussed in *SI Appendix*.

Moreover, a recent meta-analytic investigation has provided robust evidence for the link between implicit social cognition and intergroup discrimination (44). Accordingly, the endeavor to design interventions that can create enduring change in implicit social group representations has been one of fundamental importance in social cognition research. With regard to this endeavor, by establishing a correlational (study 1) and causal (study 2) link between implicit attitudes and implicit beliefs, the present results may be used to support the conjecture that implicit attitudes could be a particularly potent representation to target in such interventions. Change achieved in implicit attitudes may, in turn, lead to changes in a wide range of implicit beliefs, even seemingly nonevaluative ones, whereas change in implicit beliefs may remain relatively constrained to the specific belief targeted. We hope that future work will more systematically investigate this idea.

Finally, the present project has produced concordance in results obtained using the IAT and word embeddings derived from a large corpus of online text (study 3). As such, in line with previous work relying on a similar archival approach (45), the current project provides evidence for the idea that the automatically activated group representations indexed by the IAT are not tied to the idiosyncrasies of this particular method. Specifically, mirroring IAT results, the textual association of social group labels with positive and negative words (valence) was found to be highly predictive of social group associations with contentful semantic dimensions, such as warmth and competence (17). Taken together, these results suggest that attitude–belief dissociations occasionally obtained with explicit measures reflect the operation of deliberate cognitive processes. Specifically, when asked to characterize social groups on explicit measures of stereotype, participants may feel compelled to strategically balance negative traits with positive traits to avoid appearing prejudiced (67). However, the same processes may not be operational under more automatic conditions, including when individuals take the IAT or spontaneously produce text online. Moreover, echoing experimental evidence on the superior effectiveness of verbal manipulations in shifting implicit attitudes compared with direct experience (57), the present results raise the intriguing possibility, deserving of further empirical scrutiny, that by intervening on the kind of language that we encounter, we may be able to modulate the contents of implicit social group representations created in human minds. [We do not wish to argue that the opposite causal relationship (i.e., implicit attitudes influencing language) does not exist. In fact, we believe that the relationship between implicit social cognition and language is bidirectional. However, we highlight this particular direction here because, unlike the opposite direction, it (*i*) offers a clear point of causal intervention in the system and (*ii*) has been demonstrated empirically (57).]

It is our hope that future investigations will take up the task of providing additional evidence on the issues addressed in the present project. Specifically, even though the results obtained using word embeddings (study 3) suggest that the finding of consistent association between implicit attitudes and implicit beliefs is not restricted to the IAT, future work may probe whether the same pattern of results would be obtained using other more traditional implicit measures not involving stimulus categorization, such as sequential priming (25) or the Affect Misattribution Procedure (68). In a related vein, if our conjecture is accurate and automatically activated information tends to be inherently evaluative, then speeded self-report measures should also reveal more consistency between attitudes and beliefs than the same measures administered in a self-paced manner (39).

Moreover, the relationship between implicit attitudes and beliefs may be further probed via their relationship with third variables beyond group membership (studies 1A and 1B) and manipulations of evaluative information (study 2). Specifically, a number of individual studies seem to suggest that (valenced) implicit beliefs have incremental predictive validity over and above implicit attitudes in the context of intergroup behavior (49, 50). If the same result were consistently confirmed in more systematic investigations, such a finding would constrain findings of attitude–belief redundancy obtained in the current studies 1A–1B. In addition, the present findings of consistent attitude–belief association raise the question of whether and to what extent implicit beliefs may account for responding on measures of implicit attitude. Specifically, implicit attitudes toward lower status groups among white Americans may be primarily associated with different sets of implicit beliefs depending on the target group. For instance, negative implicit attitudes toward African Americans may be accounted for by associations with danger (69), whereas negative implicit attitudes toward Asian Americans may be accounted for by associations with lack of emotion or social skill (14). Finally, the current studies 2A–2B have provided evidence for the idea that purely evaluative information is capable of driving responding on the IAT in the absence of pertinent semantic information regarding the dimension of interest. However, the relative effectiveness of evaluative vs. semantic information in driving implicit beliefs remains an open empirical question to be investigated in future work.

Conclusion

In their pioneering work on word meaning, Osgood et al. (6) observed that "[. . .] every point in semantic space has an evaluative component [. . .], and, therefore, every concept must involve an attitudinal component as part of its total meaning." The present project provides evidence in favor of this general idea in the domain of implicit intergroup cognition, demonstrating that implicit attitudes (automatic attributions of valence to social groups) and implicit beliefs (automatic attributions of specific traits to social groups) are robustly and consistently congruent with each other. Addressing the relationship between feeling and thinking, Zajonc (35) famously noted that preferences need no inferences. The current work suggests that when it comes to automatic responding to social groups, inferences may not need much else beyond preferences.

Materials and Methods

Institutional Approval and Informed Consent. Studies 1 and 2 were granted ethical approval by the Committee on the Use of Human Subjects at Harvard University. Participants provided informed consent at the beginning of these studies. Study 3 used publicly available archival data and, as such, was exempt from institutional review.

Participants and Statistical Power. Participants in studies 1 and 2 were American adult volunteers recruited from the Project Implicit educational website (implicit.harvard.edu). Exclusion criteria are reported in *SI Appendix*. Statistical power to detect attitude–belief relationships was excellent: On average, studies had sufficient power to detect the small effect of $r_{\text{min}} = 0.24$, and obtained effect sizes (r_{obt}) exceeded r_{min} by a mean of $r_{\text{diff}} = +0.12$. In other words, studies were adequately powered to find even small correlations, whereas actually obtained correlations tended to be medium-sized, thus diminishing the possibility of type II errors (details are provided in *SI Appendix*).

Implicit Measures. In studies 1 and 2, participants completed two standard five-block IATs (27) in counterbalanced order: an attitude IAT, implemented to provide a measure of generalized group evaluation without specific semantic content (i.e., good–bad), and a belief IAT, implemented to provide a measure of specific semantic beliefs (e.g., smart–dumb, street-smart–book-smart, American–foreign). The order of critical blocks was independently counterbalanced within each IAT. Procedural details and stimuli are reported in *SI Appendix*. Performance on the IAT was evaluated using the improved scoring algorithm (70).

Explicit Measures. In studies 1 and 2, following the implicit measures, self-report measures of attitudes and beliefs regarding the target groups were also administered; however, given the focus of the present project on implicit social cognition, results obtained using explicit measures are reported only in *SI Appendix*.

Procedure. In study 1, participants' implicit attitudes and implicit beliefs were measured at baseline. The order of the two IATs was counterbalanced. By contrast, study 2 consisted of two phases: (*i*) a learning phase in which participants were randomly assigned to an experimental condition involving attitude induction via evaluative conditioning (i.e., pairing group members with valenced images) or to a control condition involving the same number of stimulus presentations without attitude induction and (*ii*) a test phase involving measurement of implicit attitudes and beliefs. In study 2A, the order of attitude and belief IATs was counterbalanced; in study 2B, the belief IAT was always administered first to prevent contamination. Details of the procedure used in study 2 are provided in *SI Appendix*.

Statistical Analyses. All statistical analyses were conducted in the R statistical computing environment. The R code for all analyses, data files (including trial-level IAT data), and materials are freely available from the Open Science Framework (71).

ACKNOWLEDGMENTS. We thank Sarah Ryan for assistance with study coding and Mark Thornton for statistical advice. This work was supported by the Dean's Competitive Fund for Promising Scholarship of the Harvard Faculty of Arts and Sciences and an Undergraduate Research Scholar award from the Institute of Quantitative Social Science at Harvard University (both to M.R.B.).

1. Heiphetz L, Spelke ES, Harris PL, Banaji MR (2013) The development of reasoning about beliefs: Fact, preference, and ideology. *J Exp Soc Psychol* 49:559–565.
2. Allport GW (1935) Attitudes. *A Handbook of Social Psychology*, ed Murchison C (Clark Univ Press, Worcester, MA), pp 798–844.
3. Eagly AH, Chaiken S (1998) Attitude structure and function. *The Handbook of Social Psychology*, eds Gilbert DT, Fiske ST, Lindzey G (Random House, Boston), pp 262–322.
4. Duncan S, Barrett LF (2007) Affect is a form of cognition: A neurobiological analysis. *Cogn Emotion* 21:1184–1211.
5. Madva A, Brownstein M (2018) Stereotypes, prejudice, and the taxonomy of the implicit social mind. *Noûs* 52:611–644.
6. Osgood CE, Suci GJ, Tannenbaum PH (1957) *The Measurement of Meaning* (Univ of Illinois Press, Urbana, IL).
7. Ghavami N, Peplau LA (2012) An intersectional analysis of gender and ethnic stereotypes. *Psychol Women Q* 37:113–127.
8. Warriner AB, Kuperman V, Brysbaert M (2013) Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav Res Methods* 45:1191–1207.
9. Katz D, Braly K (1933) Racial stereotypes of one hundred college students. *J Abnorm Soc Psychol* 28:280–290.
10. Brigham JC (1971) Racial stereotypes, attitudes, and evaluations of and behavioral intentions toward Negroes and Whites. *Sociometry* 34:360–22.
11. Brigham JC (1972) Racial stereotypes: Measurement variables and the stereotype–attitude relationship. *J Appl Soc Psychol* 2:63–76.
12. Smith AJ, Clark RD, 3rd (1973) The relationship between attitudes and beliefs. *J Pers Soc Psychol* 26:321–326.
13. Crandall CS, Bahns AJ, Warner R, Schaller M (2011) Stereotypes as justifications of prejudice. *Pers Soc Psychol Bull* 37:1488–1498.
14. Lin MH, Kwan VS, Cheung A, Fiske ST (2005) Stereotype content model explains prejudice for an envied outgroup: Scale of anti-Asian American stereotypes. *Pers Soc Psychol Bull* 31:34–47.
15. Cuddy AJC, Fiske ST (2004) Doddering but dear: Process, content, and function in stereotyping of older persons. *Ageism: Stereotyping and Prejudice Against Older Persons*, ed Nelson TD (MIT Press, Cambridge, MA), pp 3–26.
16. Lalonde RN, Gardner RC (1989) An intergroup perspective on stereotype organization and processing. *Br J Soc Psychol* 28:289–303.
17. Fiske ST, Cuddy AJC, Glick P, Xu J (2002) A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *J Pers Soc Psychol* 82:878–902.
18. Greenwald AG, Pettigrew TF (2014) With malice toward none and charity for some: Ingroup favoritism enables discrimination. *Am Psychol* 69:669–684.
19. Cuddy AJC, et al. (2009) Stereotype content model across cultures: Towards universal similarities and some differences. *Br J Soc Psychol* 48:1–33.
20. Gawronski B, Bodenhausen GV (2006) Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychol Bull* 132:692–731.
21. Rydell RJ, McConnell AR (2006) Understanding implicit and explicit attitude change: A systems of reasoning analysis. *J Pers Soc Psychol* 91:995–1008.
22. Strack F, Deutsch R (2004) Reflective and impulsive determinants of social behavior. *Pers Soc Psychol Rev* 8:220–247.
23. Bargh JA (1989) Conditional automaticity: Varieties of automatic influence in social perception and cognition. *Unintended Thought*, eds Uleman JS, Bargh JA (Guilford Press, New York), pp 3–51.
24. Devine PG (1989) Stereotypes and prejudice: Their automatic and controlled components. *J Pers Soc Psychol* 56:5–18.
25. Fazio RH, Sanbonmatsu DM, Powell MC, Kardes FR (1986) On the automatic activation of attitudes. *J Pers Soc Psychol* 50:229–238.
26. Greenwald AG, Banaji MR (1995) Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychol Rev* 102:4–27.
27. Greenwald AG, McGhee DE, Schwartz JLK (1998) Measuring individual differences in implicit cognition: The Implicit Association Test. *J Pers Soc Psychol* 74:1464–1480.
28. De Houwer J (2014) A propositional model of implicit evaluation. *Soc Personal Psychol Compass* 8:342–353.
29. Amodio DM, Devine PG (2006) Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *J Pers Soc Psychol* 91:652–661.
30. Carlsson R, Björklund F (2010) Implicit stereotype content: Mixed stereotypes can be measured with the Implicit Association Test. *Soc Psychol (Gott)* 41:213–222.
31. Rohmer O, Louvet E (2012) Implicit measures of the stereotype content associated with disability. *Br J Soc Psychol* 51:732–740.
32. Schwartz MB, Vartanian LR, Nosek BA, Brownell KD (2006) The influence of one's own body weight on implicit and explicit anti-fat bias. *Obesity (Silver Spring)* 14:440–447.
33. Bargh JA, Chaiken S, Govender R, Pratto F (1992) The generality of the automatic attitude activation effect. *J Pers Soc Psychol* 62:893–912.
34. Kahneman D (2003) A perspective on judgment and choice: Mapping bounded rationality. *Am Psychol* 58:697–720.
35. Zajonc RB (1980) Feeling and thinking: Preferences need no inferences. *Am Psychol* 35:151–175.
36. Lai CK, et al. (2014) Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *J Exp Psychol Gen* 143:1765–1785.
37. Fazio RH (2007) Attitudes as object–evaluation associations of varying strength. *Soc Cogn* 25:603–637.
38. Todd AR, Simpson AJ, Thiem KC, Neel R (2016) The generalization of implicit racial bias to young black boys: Automatic stereotyping or automatic prejudice? *Soc Cogn* 34:306–323.
39. Ranganath KA, Smith CT, Nosek BA (2008) Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *J Exp Soc Psychol* 44:386–396.
40. Smith CT, Nosek BA (2011) Affective focus increases the concordance between implicit and explicit attitudes. *Soc Psychol (Gott)* 42:300–313.
41. Gawronski B, LeBel EP (2008) Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *J Exp Soc Psychol* 44:1355–1361.
42. Zitelny H, Shalom M, Bar-Anan Y (2017) What is the implicit gender-science stereotype? Exploring correlations between the gender-science IAT and self-report measures. *Soc Psychol Personal Sci* 8:719–735.
43. Kurdi B, Gershman SJ, Banaji MR (2019) Model-free and model-based learning processes in the updating of explicit and implicit evaluations. Available at https://osf.io/f8pg3/. Accessed February 8, 2019.
44. Kurdi B, et al. (December 13, 2018) Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *Am Psychol*, 10.1037/amp0000364.
45. Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356:183–186.
46. Garg N, Schiebinger L, Jurafsky D, Zou J (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci USA* 115:E3635–E3644.
47. Esses VM, Haddock G, Zanna MP (1993) Values, stereotypes, and emotions as determinants of intergroup attitudes. *Affect, Cognition, and Stereotyping: Interactive Processes in Group Perception*, eds Mackie DM, Hamilton DL (Academic Press, San Diego), pp 137–166.
48. Gardner RC, Lalonde RN, Nero AM, Young MY (1988) Ethnic stereotypes: Implications of measurement strategy. *Soc Cogn* 6:40–60.
49. Rudman LA, Ashmore RD (2007) Discrimination and the Implicit Association Test. *Group Process Intergroup Relat* 10:359–372.
50. Hehman E, Flake JK, Calanchini J (2017) Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Soc Psychol Personal Sci* 9:393–401.
51. Tajfel H (1982) Social psychology of intergroup relations. *Annu Rev Psychol* 33:1–39.
52. Nosek BA, et al. (2007) Pervasiveness and correlates of implicit attitudes and stereotypes. *Eur Rev Soc Psychol* 18:36–88.
53. Melnikoff DE, Bailey AH (2018) Preferences for moral vs. immoral traits in others are conditional. *Proc Natl Acad Sci USA* 115:E592–E600.
54. Gawronski B, Deutsch R, Mbirkou S, Seibt B, Strack F (2008) When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *J Exp Soc Psychol* 44:370–377.
55. Glaser J, Banaji MR (1999) When fair is foul and foul is fair: Reverse priming in automatic evaluation. *J Pers Soc Psychol* 77:669–687.
56. Levey AB, Martin I (1975) Classical conditioning of human 'evaluative' responses. *Behav Res Ther* 13:221–226.
57. Kurdi B, Banaji MR (2017) Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *J Exp Psychol Gen* 146:194–213.
58. DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. *Control Clin Trials* 7: 177–188.
59. Mierke J, Klauer KC (2003) Method-specific variance in the Implicit Association Test. *J Pers Soc Psychol* 85:1180–1192.
60. Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A (2017) Advances in pre-training distributed word representations. arXiv:1712.09405. Preprint, posted December 26, 2017.
61. Pennington J, Socher R, Manning CD (2014) GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, eds Moschitti A, Pang B (Association for Computational Linguistics, Doha, Qatar), pp 1532–1543.
62. Fiske ST (2010) Envy up, scorn down: How comparison divides us. *Am Psychol* 65: 698–706.
63. Roessel J, Schoel C, Stahlberg D (2017) What's in an accent? General spontaneous biases against nonnative accents: An investigation with conceptual and auditory IATs. *Eur J Soc Psychol* 48:535–550.
64. Urdan T, Munoz C (2011) Multiple contexts, multiple methods: A study of academic and cultural identity among children of immigrant parents. *Eur J Psychol Educ* 27: 247–265.
65. Rudolph A, Hilbert A (2015) A novel measure to assess self-discrimination in binge-eating disorder and obesity. *Int J Obes* 39:368–370.
66. Levinson JD, Cai H, Young DM (2010) Guilty by implicit racial bias: The guilty/not guilty Implicit Association Test. *Ohio State J Crim Law* 8:187–208.
67. Bergsieker HB, Leslie LM, Constantine VS, Fiske ST (2012) Stereotyping by omission: Eliminate the negative, accentuate the positive. *J Pers Soc Psychol* 102:1214–1238.
68. Payne BK, Cheng CM, Govorun O, Stewart BD (2005) An inkblot for attitudes: Affect misattribution as implicit measurement. *J Pers Soc Psychol* 89:277–293.
69. Glaser J, Knowles ED (2008) Implicit motivation to control prejudice. *J Exp Soc Psychol* 44:164–172.
70. Greenwald AG, Nosek BA, Banaji MR (2003) Understanding and using the Implicit Association Test I. An improved scoring algorithm. *J Pers Soc Psychol* 85:197–216.
71. Kurdi B, Mann TC, Charlesworth TES, Banaji MR (2019) Data from "The relationship between implicit intergroup attitudes and beliefs." Open Science Framework. Available at https://osf.io/xyhgu/. Deposited January 16, 2019.