Title

Implicit Attitudes Evoked by a Singular American Slur: Five Experiments on N***er and N***a in Samples of Black and White Americans

Journal

Social Cognition, 42(3)

ISSN

0278-016X

Authors

Hudson, Sa-kiera Tiarra Jolynn

Kurdi, Benedek

Lai, Calvin K

et al.

Publication Date

2024-06-01

DOI

10.1521/soco.2024.42.3.161

Implicit attitudes evoked by a singular American slur:

Five experiments on *n***er* and *n***a* in samples of Black and White Americans

Sa-kiera Tiarra Jolynn Hudson[1*], Benedek Kurdi[2*], Calvin K. Lai[3],

Julian Johnson[4], and Mahzarin R. Banaji[5]

[1] Haas School of Business, University of California, Berkeley, Berkeley, CA

[2] Department of Psychology, University of Illinois Urbana–Champaign, Champaign, IL

[3] Department of Psychological & Brain Sciences, Washington University at St. Louis, St. Louis, MO

[4] Bushkill, PA

[5] Department of Psychology, Harvard University, Cambridge, MA

Author Note

B.K. and C.K.L. are members of the Scientific Advisory Board of Project Implicit, a 501(c)(3) non-profit organization and international collaborative of researchers who are interested in implicit social cognition.

Materials, data, and analysis scripts are available via the Open Science Framework (https://osf.io/ezhg3/).

S.T.J.H., B.K., C.K.L., J.J., and M.R.B. designed research; S.T.J.H. and B.K. performed research; S.T.J.H. and B.K. analyzed data; and S. T. J. H., B.K., and M.R.B. wrote the paper with input from C.K.L. and J.J. *S.T.J.H. and B.K. contributed equally to this work and share first authorship.

Correspondence concerning this article should be addressed to Benedek Kurdi, Department of Psychology, University of Illinois Urbana–Champaign, 603 E Daniel St, Champaign, IL 61820, email: kurdi@illinois.edu, or Mahzarin R. Banaji, Department of Psychology, Harvard University, 33 Kirkland St, Cambridge, MA 02138, email: mahzarin_banaji@harvard.edu.

Abstract

Five studies examined implicit (IAT) attitudes toward the slurs *n***er* and *n***a* among Black and White Americans (total $N = 3{,}226$). Both groups showed strong implicit negativity toward *n***er*/*a* combined relative to socially acceptable contrast terms such as *Black* or *African American*. Controlling for baseline Black–White race attitudes, Black Americans engaged in conscious reappropriation exhibited similar implicit negativity toward *n***er*/*a* to White Americans. When the rhotic and non-rhotic forms were directly contrasted, *n***er* was more implicitly negative than *n***a*, with Black Americans distinguishing the two more strongly than did White Americans. However, even Black American reappropriators showed implicit negativity toward *n***a* relative to *Black*. In sum, both *n***er* and *n***a* evoke automatic negative meaning in a broad sample of Americans today. At the same time, the relatively more positive meaning of *n***a* over *n***er* demonstrates the power of reappropriation to wrest control of word meaning.

Keywords: Implicit Association Test (IAT); implicit race attitudes; *n***er*; *n***a*; reappropriation

Implicit attitudes evoked by a singular American slur:

Five experiments on *n***er* and *n***a* in samples of Black and White Americans

*N***er*[1] is "the paradigmatic racial slur" (Kennedy, 2002, as cited in Garcia, 2003). The word goes off like a gunshot as its two syllables resurrect a dark period in American history and its legacy today. Americans, even those with little knowledge of history, are aware of the connotations of the word: to disrespect, to demean, and to instill fear through domination. In acknowledgement of the exceptional status of American slavery, the use of *n***er* is taboo in both speech and writing. If *n***er* carries the weight of slavery and dehumanization, the meaning of the variation *n***a* is more complex. *N***a*, the non-rhotic form, has been reappropriated for

---

[1] Any thoughtful scholar or scientist writing about the term *n***er* today must consider the full force of the history of the term and what it means to evoke it, even if the word is mentioned rather than used (Devitt & Sterelny, 1999) in the service of scientific communication. We considered several alternative ways of symbolizing the term in writing, and after extensive discussions opted in favor of indirect mention with replacement of several letters with a symbol, as in *n***er* and *n***a*. (The sole exceptions are direct citations and titles of relevant previous publications.) We did consider rendering the slurs exactly as participants saw and heard them in the experimental sessions (i.e., without any suppression or modification), believing that such mention would serve the function of accurate representation. However, we instead opted to use *n***er* and *n***a* to allow focus on the content of the work and avoid any distractions that may result from reporting the exact tokens that were used. The unredacted written and auditory materials to which participants were exposed are available for download from the Open Science Framework (https://osf.io/ezhg3/).

Our discussions centered on the responsibility that we carry when engaging with these words in the scientific context of exploration and discovery. We did so explicitly, given our knowledge of experimental psychology's history of racism in its theories and experimental protocols (Roberts et al., 2020). We considered our own identities, especially our racial/ethnic identities and our group origins in Africa, Europe, and Central, East, and South Asia. We reminded ourselves of our good fortune of being able to collaborate across our own diverse racial and ethnic identities because we reside in the United States — a country of immigrants, both voluntary and forced — and the power of empirical approaches to difficult social problems. We reflected on our duty, as students of social cognition and society, to be mindful of the impact of our research in the context of societal norms today. We also reflected, to the extent our imagination permitted, on the ways in which this work will be viewed in the future.

use by segments of Black American culture, both for self-reference and in reference to Black American friends and loved ones. This type of use is affiliative and represents an attempt toward conscious reappropriation.

It is rare that a single word from a language can raise the range of questions of both scientific and societal interest that *n***er* and *n***a* do. Yet, no empirical study of their attitudinal status, explicit or implicit, currently exists. In this paper we remedy this situation by providing the first tests of self-reported (explicit) and automatic (implicit) attitudes evoked by the slurs, jointly and separately, in large samples of Black and White Americans. Although not of primary interest, we include measures of explicit attitudes, which we expect will reflect conscious awareness of American history as well as the taboo status of the slur. The more pertinent question for psychological theory about evaluative representations of the term centers on implicit attitudes, or evaluations that are revealed on measures of relatively more automatic thought and feeling[2] (Devine, 1989; Fazio et al., 1986; Greenwald & Banaji, 1995).

Implicit attitudes are commonly assumed to reflect the sum total of the evaluative associations in the context of which a concept has been encountered. As such, based on standard associative accounts of implicit social cognition (McConnell & Rydell, 2014; Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000), along with basic tenets of reinforcement theory (Skinner, 1958), the disuse of *n***er* in spoken and written language should have caused its evaluative potency to dissipate. After all, without reinforcement, the negativity associated with a stimulus

[2] Throughout this paper, we use the term "attitude" to refer to evaluative representations stored in long-term memory (e.g., Fazio, 2007). We use the term "implicit attitudes" to refer to attitudes retrieved under relatively automatic (suboptimal) conditions, such as in the absence of the person's intention, awareness, or control (De Houwer et al., 2009). We contrast implicit attitudes with explicit attitudes, which are retrieved under relatively controlled (optimal) conditions. Although no one-to-one correspondence between measures and underlying processes can be assumed (Jacoby, 1991), implicit attitudes are customarily indexed using indirect measures, such as the Implicit Association Test (IAT; Greenwald et al., 1998), whereas explicit attitudes are customarily indexed using direct measures of self-report.

should fade over time. As such, in populations that do not use the term and rarely encounter it, implicit attitudes evoked by it may have neutralized through disuse.

However, reinforcement through usage is not the only way in which learning is established and maintained in humans. Specifically, in this case, abstract propositional knowledge of the origins and taboo status of the term may be sufficient to have maintained the negative valence associated with the it in spite of disuse and lack of reinforcement. Evidence to support this idea comes from research showing that implicit attitudes are sensitive not only to incremental associative learning; rather, a single propositional statement can create implicit attitudes that are even more robust than those created by many repeated pairings (De Houwer, 2014; Kurdi & Banaji, 2017; Kurdi & Dunham, 2020).

Parallel work on the communicative role of punishment suggests that societal sanctions around *n***er* may lead to stability in negative implicit attitudes, regardless of levels of usage in society (Sarin et al., 2021). Specifically, according to this account, the primary role of punishment is to alter behavior durably without the need for constant reinforcement. As such, once knowledge of its taboo status is acquired, both explicit and implicit attitudes toward the slur may demonstrate negativity — even in the long term and without repeated negative consequences of use. The data from White Americans in particular (a group of theoretical interest given that it has experienced widespread disuse of *n***er* for decades) will be germane to the issue of how evaluative (positive/negative) meaning is maintained over time, especially as it is revealed automatically.

Evaluative responses measured among Black Americans will help answer another, related question. Although *n***er* is consensually distinctly negative in American culture, evaluative representations of the non-rhotic variant, *n***a*, may be more complex. Its frequent use by Black

artists in hip-hop, rap, improv, and stand-up comedy has made it widespread in those subcultures, especially among young adults (Tesler, 2015). In fact, there is likely no other English word that is as taboo in the larger culture and as profusely used in a subculture as *n***a*. Although the initial studies reported below combined the two words to probe their joint meaning, testing the difference between the attitudes evoked by the rhotic and non-rhotic versions is theoretically relevant and the focus of the later studies. Specifically, if the two terms are, indeed, evaluatively distinct from each other, then such differences should be apparent not only on measures of explicit attitudes but even on measures of implicit attitudes.

In considering potential patterns of data that may emerge following reappropriation, we revisit research conducted in the 1990s to study the implicit stereotypes evoked by gendered words in language, such as *chairman*, *actress*, and *guys* (Banaji & Hardin, 1996). Many at the time believed that although terms of this kind contain explicitly gendered morphemes (such as "man"), they ought to be treated as generic because the user's intention was to be inclusive. The morphology of the word was thought to be irrelevant, with the male generic used merely in the service of efficiency.

However, all 150 male generics tested in an experiment relying on a well-established semantic priming procedure (Meyer & Schvaneveldt, 1971; Neely, 1976) were found to activate the concept MALE (and not FEMALE), no matter the communicator's intent (Banaji & Hardin, 1996). The sole exception to this finding (among college undergraduates) was the term *freshman*, which activated FEMALE and MALE equally, presumably because it had been used to refer to both female and male students equally often over a relatively long period of time. This latter result points to the possibility that, given protracted use that differs markedly from its highly negative

original connotation, the automatically evoked meaning of *n***a* may have shifted toward positivity.

A growing body of research on the origins of implicit social cognition in processes of learning and memory further suggests that although encountering targets in positive or negative contexts is a powerful way to establish and maintain implicit attitudes toward them, implicit attitudes can also shift as a result of effortful rejection of experienced valence (for a review, see Kurdi et al., 2022). Thus, the non-rhotic form, *n***a*, devised as a specific variant to combat the negativity of the original *n***er* may have allowed new, positive meaning to emerge — at least among those who use it in affiliative contexts. In line with this possibility, recent evidence also points to sizeable societal-level changes in implicit sexuality, race, and skin tone attitudes over the course of 14 years (Charlesworth & Banaji, 2022a). Given these considerations, the time is ripe to ask whether usage of these slurs in communities that employ them in affiliative contexts has changed their meaning sufficiently to make them positive (or at least noticeably less negative).

***

So far, we have remained close to the scientific questions about evaluative learning and representation that motivated this research. However, discussions about the terms, in society and the law, are very much alive in present-day United States, and the results of the studies reported here can inform such debate. Measures of implicit attitudes are critical given that, relatively speaking, they bypass intentional control to assess the basic semantic and affective meaning of attitude targets, including — in this case — particular words. Cultural debates surrounding the non-rhotic version (*n***a*) in the public sphere are not currently informed by empirical evidence about its cognitive–affective status, given that no such evidence is available.

There is a sense in contemporary American society that Black Americans are free to use the slurs but that such use is off limits to White Americans and other non-Black Americans (Kirkland & Silverman, 2018). Indeed, use of the terms by White Americans regularly produces criticism and censure, and it can even result in the loss of a job (e.g., the case of Donald McNeil, long-time reporter at *The New York Times*, who was fired for use of the slur on a trip with students; Montgomery & Cartwright, 2021). However, to demonstrate the complexity of relevant attitudes, we note that there are several videos on social media depicting White students singing lyrics to rap songs that include *n***a*. At the same time, such usage has often received media attention and resulted in punitive actions (Bauer-Wolf, 2018, 2019). Indeed, questions about who can use the term toward whom and where (in what context) are contested, and individual cases sprout up regularly in the media.

Opposing views about the use of the slurs also continue to pose challenges to institutions that must adjudicate conflicts that emerge from intentional or accidental usage of the slurs. Indeed, disagreement surrounding the use of *n***a* among Black Americans themselves has reached the federal courts. At least two cases (*Johnson v. Strive*; *Weatherly v. Alabama State University*) have centered on intragroup use of *n***a* by Black Americans in supervisor/employee relationships. In both cases, the court has held that use of the slur by Black supervisors directed at Black employees did not remove the racial sting associated with the word and therefore constituted racial harassment.

Debates about the meaning and use of *n***a* abound in Black American culture itself. Older Black Americans and long-standing institutions such as the National Association for the Advancement of Colored People (NAACP) explicitly consider its usage by Black Americans to be harmful to the community and do not favor usage. Others, such as the writer Ta-Nehisi

Coates, defend the use of the words by Black Americans for Black Americans (Coates, 2013), and others agree with his nuanced approach that considers the identities of both speaker and listener (Kennedy, 1999; McWhorter, 2010; Rahman, 2012). As of now, as deep as these personal convictions might be, adjudicating between opposing views cannot be guided by evidence if the cognitive–affective status of these words remains unknown.

Given questions of scientific interest concerning reinforcement versus propositional knowledge as a source of automatically revealed word meaning and the fraught questions of cultural interest discussed above, what does psychological science have to say about the cognitive–affective status of the slurs? The experiments reported below are the first ones to undertake an empirical investigation of the mental status of these slurs. These are likely to be the first tests because — as we ourselves have experienced — there is discomfort in approaching the topic at all. However, as with any topic that creates discomfort, thoughtful tests and careful inferences made from such tests can do a lot to unveil ignorance. Indeed, the studies reported below inform about the standing of the slurs in the minds of a large and diverse group of Americans in the third decade of the 21st century. With all materials, data, and analysis scripts available in an open repository, these data can serve as a baseline to which the data from future generations may be compared to produce evidence of attitude stability and change.

The Present Project

Driven by the considerations laid out above, the goal of the present project was to probe the attitudinal standing of the words *n***er* and *n***a* in contemporary samples of Black and White Americans, with primary focus on implicit (automatic) attitudes. A high-level overview of the design of each study is provided in Table 1.

In Study 1, we introduce a new semi-auditory Implicit Association Test (IAT) and use it to measure implicit attitudes toward *n***er*/*a* combined in a racially diverse sample of adult American volunteers. The goals of this study were twofold: First, we sought to validate the modified semi-auditory IAT procedure, to be used in this and all subsequent studies, in which category stimuli (such as *n***er* and *Black*) were presented auditorily and valenced attribute stimuli (such as *angel* and *awful*) were presented visually, as on regular IATs. If this partial auditory procedure is found to be viable, that result would also contribute to the development of indirect measures that could be used among visually impaired individuals. At present, no other indirect tasks allow for auditory testing, thus leaving visually impaired individuals out of the database of implicit social cognition research. The second aim of this study was to provide the first empirical evidence on the evaluative status of the slurs relative to socially acceptable contrast terms, both with a different referent (specifically, *White* and *European American*) and the same referent (*Black* and *African American*).

In Study 2, both implicit race (White–Black) attitudes and implicit attitudes evoked by the slurs combined were measured in two samples of theoretical interest, allowing us to capture a snapshot of the reappropriation process unfolding. A contribution of this study, perhaps the theoretically most critical one, will be the opportunity to answer the question: Are implicit attitudes toward the slurs more positive in a sample of young (mostly Black and Hispanic) reappropriators than they are in an online sample of volunteers (mostly White non-reappropriators)? Both possibilities are conceivable. It is possible that reappropriation of the term in expressly affiliative contexts has changed the automatic meaning of the term to be positive in valence. Alternatively, knowledge of the history of the term and the current widespread cultural taboo on its use might render the term to be negative in meaning even among a subculture of reappropriators, their

intentions notwithstanding. As explained below, in this study, the White–Black race IAT (independent of the slur IATs) served as a critical baseline test to which to anchor group differences in implicit attitudes toward the terms.

Combining the rhotic and non-rhotic versions of the term seemed defensible for the initial tests conducted in Studies 1–2. However, as explained above, *n***er* and *n***a* differ in both their history and status in contemporary American culture, especially among Black Americans, with the former virtually universally seen as taboo and the latter used by some segments of Black American society with the conscious goal of reappropriation. As such, Study 3 contrasted the rhotic (*n***er*) and non-rhotic (*n***a*) versions of the term directly with each other to probe whether the culturally relevant distinction between the two is sufficiently potent to permeate even automatic aspects of cognition and affect. We did so in a numerically balanced online sample of Black and White Americans, driven by the recognition that both frequency of exposure and perceived acceptability of usage is expected to differ across these two populations. As such, along with Study 2, the contribution of this study is the additional insight that it will provide into the reappropriation process.

Finally, Studies 4A and 4B contrasted each term, *n***er* and *n***a*, separately with *Black*, in a racially diverse sample of adult Americans (Study 4A) and in a numerically balanced sample of Black and White Americans (Study 4B). These studies provide a more conservative test of the questions addressed in Study 3 given that they probe the difference between the rhotic and non-rhotic versions relative to a widely accepted contrast term with the same referent. In addition, as a further exploratory test of generalizability, all three final studies featured a manipulation of the modality in which participants were exposed to the slurs. Similar to Studies 1–2, the auditory conditions of Studies 3, 4A, and 4B relied on presentations of the slurs in the auditory

modality, whereas the newly introduced visual conditions featured visual stimulus presentations, similar to most regular IATs.

As explained above, the present studies also speak to the general theoretical question of the origins of implicit attitudes in environmental contingencies and human information processing. Specifically, dominant associative perspectives on implicit social cognition predict that the slurs should have lost their evaluative potency through disuse, especially when it comes to the rhotic version and to samples of White Americans. Any evidence that the terms are still associated with automatic negative valence in these participants' minds would point to the possibility that implicit attitudes emerge from mechanisms other than rote association formation, including — potentially — propositional knowledge of social norms and the taboo status of the terms.

Together, these five studies also provide tests of generalizability of the central findings along multiple axes, including the sample of interest (White non-reappropriators, Black non-reappropriators, and Black reappropriators), version of the word (rhotic vs. non-rhotic), contrast terms (direct comparison vs. indirect comparison to socially acceptable terms, such as *Black* or *African American*), the modality of stimulus presentation (auditory vs. visual), and the identity (gender and race) of the speaker. At the same time, as we explain in more detail in the General Discussion, we hope that future work will expand on the present project in at least two ways: (*a*) by considering the use of measures beyond the IAT, potentially including those that can capture (implicit) ambivalence and (*b*) by more directly exploring contextual modulations in these effects, as a function of features of the situation such as the linguistic context, the physical environment, as well as the identity of both the speaker and their interaction partner.

Study 1

The aim of Study 1 was twofold: first, we sought to validate the new semi-auditory Implicit Association Test (IAT) procedure used in this and all remaining studies; second, once sufficient evidence for validity was obtained, we measured attitudes evoked by *n***er*/*a* combined[3] in a diverse online sample of adult American volunteers. The IAT was selected for use in all studies given its strong psychometric properties relative to other measures of implicit cognition (Bar-Anan & Nosek, 2014), along with the fact that — unlike several alternatives, such as sequential priming — it does not require hundreds of exposures to the slurs to produce reliable effects. At the same time, it should be noted that the IAT is a relatively decontextualized experimental measure of attitudes and, as such, cannot be expected to capture all subtle situational variations in the evaluative content associated with the slurs. We return to these issues in the General Discussion.

Our interest in (*a*) attitudes evoked by a particular word and a variation of it and (*b*) presenting the slur words auditorily (rather than visually) created some challenges for the standard IAT procedure, which has typically been used to study a category of which multiple exemplars are available. In fact, previous research suggests that variation in exemplars used is important for the validity of the procedure, and at least four exemplars of a category are needed to create a robust test (Greenwald et al., 2022). Moreover, standard IATs conducted with adult participants usually involve visual presentation of all stimuli.

As such, a subset of participants recruited for Study 1 were randomly assigned to complete a semi-auditory version of the standard race IAT, measuring implicit attitudes toward Black

---

[3] Throughout the paper, when the two terms were collapsed to represent a single category, we use the notation *n***er/a*. When the two terms are analyzed separately in Studies 3, 4A, and 4B (to test differences in implicit attitudes evoked by each) we retain the separation visually in labels, i.e., *n***er* vs. *n***a*.

and White Americans (referred to as set 1 below). On this semi-auditory IAT, category labels (such as *Black Americans* and *White Americans*) appeared visually, but exemplars of these categories were presented auditorily (e.g., the participant heard the words *Black* and *White*). In line with the standard procedure, good and bad attribute stimuli (such as *angel* and *awful*) were presented visually. This procedure was modeled after the semi-auditory IATs that, for decades, have been used in research with young children who cannot read (Baron & Banaji, 2006; Cvencek et al., 2011; Parise & Spence, 2012). The main purpose of the tests included in set 1 was to validate the semi-auditory procedure by comparing mean levels of implicit attitudes obtained here to the same test using the standard visual procedure in large samples of adult Americans online.

Once such evidence of validity was obtained on tests in set 1, we were able to turn to two other sets of tests that were of theoretical interest: set 2 in which we used the semi-auditory procedure to contrast *n***er*/*a* with variations of the label *White* and set 3 in which we used the semi-auditory procedure to contrast *n***er*/*a* with variations of the label *Black*. Given that the referents of the non-White category labels on the White–Black race attitude IAT (set 1) and the White–*n***er*/*a* IAT (set 2) are the same, any difference between the two sets is indicative of differences in connotation (more specifically, evaluation) of the socially acceptable term *Black* and the taboo words *n***er*/*a*. Finally, if differences in implicit attitude emerge on the Black–*n***er*/*a* IAT (set 3), they can be attributed only to the unique connotations of *n***er*/*a* rather than attitudes toward the social category itself, thus making this test a particularly stringent one.

Method

Open Science Practices

All materials, data, and analysis scripts for this and all remaining studies are available for download from the Open Science Framework (https://osf.io/ezhg3/). We report how we

determined the sample size, all data exclusions, all manipulations, and all measures in this and all remaining studies. The study designs and statistical analyses were not preregistered.

Participants

We recruited volunteers from the United States via the Project Implicit educational website (http://implicit.harvard.edu/). From the pool of participants arriving at Project Implicit, 1,022 were randomly assigned to the present study. Raw data were prepared using standard procedures (Greenwald et al., 2003). Specifically, we excluded participants from analyses if they did not complete the IAT, which served as the main dependent measure (n = 191), or had response latencies that were below 300 ms on 10 percent or more of IAT trials, indicating inattention (n = 14).

Following participant exclusions, the final dataset consisted of 817 participants. 433 participants identified as female and 377 participants as male. Most participants identified as White (n = 570), followed by self-identification as Black (n = 83), Hispanic (n = 67), and Asian (n = 33). Mean participant age was 37 years (SD = 15 years). Most participants were politically liberal (n = 490) rather than conservative (n = 101) and had completed an average of 17 years of education (SD = 2 years).

Given the socially sensitive nature of the study, we sought to ascertain that the data were not subject to different forms of attrition that would endanger the validity of statistical inferences.[4] One potential threat to validity is excessive attrition, i.e., participants leaving the study in sufficiently large numbers to raise the specter of the remaining sample not appropriately resembling the participant pool as a whole. Another potential threat to validity is selective attrition whereby members of certain demographic groups are less likely to complete the study than

[4] This suggestion of possible threat to generalizability was offered by an anonymous reviewer, for which we are grateful given that we were able to rule out such a possibility.

others, again causing distortions to sample composition. We have found neither of these two processes to be operating in the data of Study 1 or any of the remaining studies. We also did not observe any differential attrition across IAT variants in any of the studies. Further details are available in the online materials.

Procedure

In the consent phase, participants were informed that the study involved seeing and hearing derogatory words (specifically, *n***er* and *n***a*). Although participants were recruited on a voluntary basis, given the use of sensitive terms, we sought to maximize the ease of dropping out of the study by multiple mentions of the possibility of leaving without negative repercussions, including on a warning screen prior to the consent form and in the consent form itself. Indeed, among the benefits of recruiting online volunteers is the complete lack of coercion given that no compensation is offered to participants and no experimenter is physically present when participants make decisions about whether to enroll in and complete the study. Participants were randomly assigned to one of seven IATs (see below for details) before completing self-report measures of attitudes toward and beliefs about the words. Finally, participants received feedback on their IAT score and were debriefed.

Implicit Association Tests. Participants were randomly assigned to one of seven attitude Implicit Association Tests (IATs; Greenwald et al., 1998), each of which can be conceptualized as belonging to one of three larger sets. The attribute dimension on all IATs was always the same (good/bad). The category labels for the two categories (e.g., "Black" vs. "White") and the stimuli used to represent each of these categories varied.

Set 1 IATs (White–Black Tests). Set 1 served as a test of whether the semi-auditory IATs performed as expected by comparing their results to the data from purely visual IATs obtained

from over 1.5 million tests taken (Charlesworth & Banaji, 2019; Nosek et al., 2007; Ratliff et al., 2020). The two IATs in set 1 measured implicit attitudes toward White and Black Americans (i.e., the standard race IAT). The only difference between the two IATs in set 1 was the label used to refer to Black/African Americans. "African American/Afro American" vs. "White/White American" served as category labels on IAT 1, whereas "Black/Black American" vs. "White/White American" served as category labels on IAT 2. These two variations are not expected to produce large differences in implicit race attitudes, but they provide a useful test of the semi-auditory IAT effect.

Set 2 IATs (White–Ner/a Tests).*** The two IATs in set 2 were designed to test the implicit attitudes elicited by the words *n***er* and *n***a* relative to White American as the contrast category. Specifically, IAT 3 used the category labels "White/European American" vs. "*N***er*/*N***a*" and IAT 4 used the category labels "White/White American" vs. "*N***er*/*N***a*".

Set 3 IATs (Black–Ner/a Tests).*** In this set, different versions of widely used and socially acceptable group labels referring to Black Americans served as category labels on the IAT, including "African American/Afro American" (IAT 5), "African American/Black" (IAT 6), and "Black American/Black" (IAT 7). In each case, one of these labels was contrasted with the category "*N***er*/*N***a*."

A standard IAT requires a minimum of four instantiations of a category or attribute (Greenwald et al., 2003). For example, if the attribute consists of pleasant and unpleasant words, at least four stimuli of pleasant and unpleasant meaning must be used to provide a reliable measure of the effect. Although we could do this easily with the good–bad attribute dimension, generating four variations of *n***er*/*a* was not possible. Instead, we used a variation of voices saying

the words *n***er* and *n***a*. These voices were computer-generated male voices using a standard American accent in Studies 1 and 2 (five voices in total) and the voices of Black American speakers (six male and four female) in Studies 3 and 4, as well as Supplementary Study 1 (ten voices in total).

Each IAT was a five-block attitude IAT in which participants sorted the stimuli that were presented visually or auditorily using the left ("E") or right ("I") response key. In the first block (attribute practice; 20 trials), participants sorted positively and negatively valenced words presented visually using "Good" and "Bad" as attribute labels. Positive attribute stimuli included the words "angel," "beauty," "delight," "enjoy," "flower," "lovely," "merry," "peaceful," and "success." Negative attribute stimuli included the words "awful," "bitter," "disgust," "evil," "failure," "lousy," "maggot," "poison," and "stinky." The two sets of words were matched on syllable length and first letter. In the second block (category practice; 20 trials), participants sorted auditory stimuli corresponding to one of the seven social category combinations (depending on the specific IAT). For example, on the "Black American/Black" vs. "*N***er*/*N***a*" IAT, they would use one key to sort the words "Black American" and "Black" and the other key to sort the words "*n***er*" and "*n***a*."

In the first critical block (combined block; 40 trials), attribute and category sorting were combined such that participants sorted one attribute and one category using the same response key (e.g., "Black American/Black" + "Good," "*N***er*/*N***a*" + "Bad"). In the fourth block (reversed category practice; 20 trials), the category labels were mapped onto the opposite key relative two blocks 1 and 3. Finally, in the second critical block (combined block; 40 trials) attribute and category sorting were combined again, but with the mappings between categories and attributes reversed compared with the first critical block. For example, if participants sorted

"Black American/Black" category stimuli with positive and "*N***er*/*N***a*" category stimuli with negative attribute stimuli earlier, in this block they were instructed to associate "Black American/Black" with bad and "*N***er*/*N***a*" with good.

The order of the two critical blocks, as well as the initial assignment of category and attribute stimuli to the left and right keys, was randomized. Performance on the IAT was evaluated using the improved scoring algorithm (Greenwald et al., 2003). Positive D scores in set 1 indicate implicit preference for the category White over Black, in set 2 implicit preference for the category White over the words *n***er*/*a*, and in set 3 implicit preference for the category Black over *n***er*/*a*.

Self-Report Measures. Participants completed a series of self-report measures in randomized order. Given time constraints, we administered abbreviated versions of the modern racism (3 items; McConahay, 1986), social dominance orientation (4 items; Ho et al., 2015), and racial identification scales (2 items; Sellers et al., 1998). Most likely due to the small number of items used, internal consistency of these scales was below the standard cutoff for acceptable reliability at 0.60 (Cronbach's $\alpha s < 0.57$). Thus, we do not discuss them further.

In addition, participants were instructed to report their experiences with the words *n***er* and *n***a* and their beliefs about the meaning and usage of each word. Responses were provided on 7-point Likert scales whose endpoints were labeled 1 (not at all) and 7 (very much so), respectively. For each of the two words, participants were asked (*a*) how frequently they hear the word in their everyday life (frequency of exposure); (*b*) how often they use the word (frequency of usage); (*c*) how offensive they consider the word (offensiveness–self); (*d*) to what degree the offensiveness of the word depends on the situation or who is saying the word (contextual

variation); and (*e*) whether they think there is a general consensus that the word is offensive (regardless of they feel about the word; offensiveness–society).

Results

Implicit Attitudes

Figure 1 shows the data from collapsed tests in sets 1, 2, and 3 (see online materials for a breakdown of results by individual tests in each set). Descriptively, participants showed similar implicit attitudes within each set, irrespective of the specific labels used. Overall, implicit attitudes in set 1 exhibited a preference for White over Black; implicit attitudes in set 2 exhibited a preference for White over *n***er*/*a*; and implicit attitudes in set 3 exhibited a preference for Black over *n***er*/*a*. The smallest effect was detected in set 1, a larger effect in set 2, and the largest effect in set 3.

To formally test whether (*a*) mean implicit attitudes were comparable to each other within each set and (*b*) differed from each other across the three sets of IATs, we conducted Bayesian modeling using equality and ordinal constraints imposed upon the seven IAT means (Haaf & Rouder, 2017; Rouder et al., 2018). Specifically, we set means to be equal to each other within each set and specified the magnitude of implicit preferences to be stronger in set 3 than in set 2 and stronger in set 2 than in set 1. The resulting model clearly outperformed both the null hypothesis (assuming that all IAT means are equal), BF = 10384.27, and the full model (which allows IAT means to vary in any non-specified pattern), BF = 5.87 $\in$ [0, 6].

We then investigated the magnitude of implicit preferences in each set of IATs. The mean D score of 0.26 (*SD* = 0.48) in set 1 (White–Black attitudes) was significantly different from zero and represented a medium effect, $t(249) = 8.73$, $p < .001$, Cohen's $d = 0.55$, $BF_{10} = 1.38 \times 10^{13}$. This effect size is comparable to the effect observed on the standard race IAT using

visual stimuli (mean D = 0.32; Charlesworth & Banaji, 2019), thus underscoring the soundness of the procedure used in the present study. IATs in set 2 (White–*n***er*/*a* attitudes) also produced a significant effect, which was large in size ($M = 0.36$, $SD = 0.45$), $t(227) = 11.94$, $p < .001$, Cohen's $d = 0.79$, $BF_{10} = 4.60 \times 10^{22}$. Finally, an even larger and statistically significant effect was observed on IATs in set 3 (Black–*n***er*/*a* attitudes; $M = 0.45$, $SD = 0.42$), $t(338) = 19.75$, $p < .001$, Cohen's $d = 1.07$, $BF_{10} = 5.87 \times 10^{54}$. Notably, the effect observed in set 3 was on par with the strongest known implicit social group attitudes, such as age (mean D = 0.44) or body weight (mean D = 0.40; Charlesworth & Banaji, 2019).

Secondary Analyses

Implicit attitudes toward *n***er*/*n***a* were unrelated to any of the explicit items administered in the study. Specifically, none of the five explicit items measuring experiences with and views of *n***er* was related to either set-2 IATs, $F(5, 213) = 0.96$, $p = .441$, $BF_{01} = 121.48$, or set-3 IATs, $F(5, 327) = 0.64$, $p = .667$, $BF_{01} = 566.23$. The same was true for the five explicit items measuring experiences with and views of *n***a* with respect to both set-2 IATs, $F(5, 213) = 0.95$, $p = .452$, $BF_{01} = 125.93$, and set-3 IATs, $F(5, 327) = 0.78$, $p = .565$, $BF_{01} = 415.82$. Further details, including analyses for each item separately, are available in the online materials and reinforce the same conclusions reported here.

Similarly, participant gender ($BF_{01} = 5.30$), ideology ($BF_{01} = 131.55$), and education ($BF_{01} = 6.43$) were unrelated to implicit attitudes toward *n***er*/*a*. We found a small but significant effect for age, $\beta = 0.18$, $t(564) = 4.42$, $p < .001$, $BF_{10} = 1156.13$, such that older participants exhibited more implicit negativity toward the slurs than younger participants did. Finally, although the means showed slightly greater implicit negativity toward *n***er*/*a* among White participants ($M = 0.44$, $SD = 0.43$) than among either Black participants ($M = 0.37$, $SD = 0.51$) or

participants of other races ($M = 0.34$, $SD = 0.42$), inferential evidence for any difference across the three groups remained inconclusive, $F(2, 564) = 2.78$, $p = .063$, $BF_{01} = 2.39$. As such, we refrain from interpreting this effect.

Discussion

Study 1 confirmed that the semi-auditory IAT was a viable method to study implicit attitudes toward the slurs given that the semi-auditory White–Black race IAT produced similar mean levels of implicit attitudes as the standard (purely visual) White–Black race IAT. With that reassurance, the present results produced robust evidence that the terms *n***er*/*a* (combined) elicit negative implicit attitudes. The negative valence associated with the terms was not sensitive to variations in the labels used to denote the Black and White contrast categories. Moreover, this result did not vary as a function of participant demographics other than a relatively small effect of age. (The effect of race was further explored in Studies 2–3 and 4B, which relied on balanced samples of Black and White Americans and was therefore better powered to detect race effects.)

Notably, the data from Study 1 suggest that (*a*) implicit preferences for White over *n***er*/*a* were stronger than implicit preferences for the category White over Black and (*b*) implicit preferences for Black over *n***er*/*a* were stronger than implicit preferences for White over *n***er*/*a*. Both of these findings are noteworthy. The first comparison indicates that implicit negativity toward the slurs is not reducible to implicit attitudes toward White and Black Americans; the second comparison is even more striking because it shows an implicit preference for the category Black over *n***er*/*a* although the group referent of both terms is the same (Black Americans). As such, the evaluative connotations of the slur words can be the only source of the implicit preference.

In conclusion, the slurs are characterized not only by negativity but rather exceptionally negative automatic meaning in the minds of both White and Black Americans, equivalent in strength to the strongest implicit social attitudes measured using the IAT. These results are also noteworthy in challenging a simple reinforcement learning account of how such negative meaning emerges and is maintained given that the terms have been erased from the language use of most Americans and, as such, are rarely encountered at all, let alone in the context of negative accompanying stimuli or punishments. Yet, the slurs are, even on measures of implicit attitude, robustly negative in meaning. Alternative theories of how the valence of word meaning remains active even in disuse, not relying on association formation or reinforcement, will be required to explain these results.

Study 2

The Black and White American participants recruited for Study 1 were drop-in volunteers. As such, given that they share an interest in research in general and implicit social cognition in particular, they can be expected to have other shared values. A strong test of valence transformation of the slurs would involve a population of Americans who use the term routinely and in an affiliative context that involves peer support. After all, it is unusual if not unique that a term that is taboo in the larger culture is thriving in a subculture, but that is the case with *n***er/a*, where especially young Black and Hispanic Americans are engaged in a broad-based and conscious attempt to reappropriate the slur.

As such, in Study 2, we recruited a sample of young, largely Black and Hispanic, underserved Americans known a priori to use the slurs in everyday linguistic social interaction. This study provides an opportunity to detect implicit positivity toward the term if it has indeed

transformed into one with positive valance in the automatic cognition of those who use it in a friendly, affiliative context.

The procedure of Study 2 was similar to Study 1, with one crucial difference: Instead of one IAT, each participant completed two IATs, one measuring implicit race attitudes (White–Black comparison) and one measuring implicit attitudes toward the slurs relative to Black as the contrast category. Given that we expected more positive implicit attitudes toward Black Americans among the underserved youth than on Project Implicit, we use the race IAT as a baseline to interpret the results of the *n***er/a* IAT in each sample (as we did with sets 1 and 3 in Study 1).

Method

Participants

Two samples were recruited. 253 participants were recruited from the Project Implicit website (referred to as "PI sample" below), and 159 participants were recruited through an organization providing educational and career support to youth from underserved communities in the United States (referred to as "UY sample" below). 73 participants were excluded from analyses for failing to complete either IAT (43 in the PI sample and 30 in the UY sample), and 13 participants were excluded from analyses for responding faster than 300 ms on 10% or more of the trials on both IATs (7 in the PI sample and 6 in the UY sample). Participants who had usable data from either IAT were retained but data from incomplete IATs or IATs with excessive rapid responding were discarded. As such, the total final sample consisted of 326 participants (203 in the PI sample and 123 in the UY sample).

As expected, the two samples differed markedly from each other in age and race (among other variables). Age and race are, indeed, the demographic variables that typify groups in the United States who are likely to reject usage of *n***er/a* and those explicitly involved in

reappropriation. The mean age of the PI sample was 38 years (*SD* = 13 years), while the mean age of the UY sample was 20 years (*SD* = 2 years). Whereas the PI sample consisted of 66% White, 12% Black and 10% Hispanic participants, 53% of the UY sample were Black and 37% were Hispanic, with no White students included. Finally, the PI sample skewed toward women (75% self-identified women), whereas the UY sample skewed toward men (64%). However, we note the lack of any effects of participant gender in any of the studies.

Procedure

The procedure for Study 2 was similar to Study 1, with one crucial change. Specifically, instead of one IAT, participants were asked to complete two IATs in randomized order: a White–Black race attitude IAT and a Black–*n***er*/*a* attitude IAT. After completing the two IATs, participants responded to five sets of explicit items in randomized order: (*a*) the same racial identification scale as in Study 1, (*b*) a set of items measuring the frequency of using and hearing the words, (*c*) feeling thermometer items measuring explicit attitudes toward Black and White Americans as well as toward the words *n***er* and *n***a*, (*d*) a 4-item scale regarding the acceptability of using the words *n***er* and *n***a* in their social groups, and (*e*) a longer scale, included for exploratory purposes, regarding experiences with the two words, with special regard to the context dependence of the acceptability of its usage (including the race of the speaker and the addressee, as well as the speaker's intent). The latter scale is not discussed further but the items and data are available in the open materials. Finally, the UY sample answered brief demographic items, and both samples received feedback on their IAT performance and were debriefed.

Implicit Association Tests. The IATs in Study 2 followed the same five-block semi-auditory protocol as in Study 1. The race attitude IAT used the category labels "White/White

American" and "Black/African American" and included the corresponding category stimuli and the same attribute labels and attribute stimuli as the IATs in Study 1. The *n***er*/*a* attitude IAT used "Black/African American" and "*N***er*/*N***a*" as category labels and stimuli and the same attribute labels and attribute stimuli as the IATs in Study 1. Positive D scores indicate an implicit preference for White over Black and Black over *n***er*/*a*, respectively.

Self-Report Measures. Participants completed the same two-item racial identification scale as Study 1. The internal consistency of this scale was poor (Cronbach's $\alpha = 0.58$) and is therefore not discussed further. Participants indicated how warmly or coldly they felt toward Black people, White people, and the words *n***er* and *n***a* using 10-point Likert scales. Participants additionally indicated how acceptable and offensive it is to say each of the two words in their social groups using 7-point Likert scales anchored with "strongly disagree" and "strongly agree."

Results

Preliminary Analyses

In a first set of analyses we ascertained that the two samples differed from each other in systematic and expected ways. Critically, although participants from the UY sample expressed dislike for the word *n***er* ($M = 1.93$, $SD = 1.64$), their attitudes toward *n***a* were close to the midpoint of the 10-point scale ($M = 4.88$, $SD = 2.56$), with 40 percent of participants reporting a positive explicit attitude. By contrast, the overwhelmingly White PI sample expressed negativity toward both terms to a similar degree ($M = 1.67$, $SD = 1.41$, for *n***er* and $M = 2.11$, $SD = 1.70$ for *n***a*). The difference between the two samples was significant only for *n***a* ($p < .001$, $BF_{10} = 6.40 \times 10^{18}$) but not for *n***er* ($p = .197$, $BF_{01} = 3.02$).

In addition, in line with expectations, the UY sample reported greater likelihood of hearing both versions of the slur in their daily lives than did the participants in the PI sample. The UY sample also reported relatively high levels of acceptability of using both words (and especially *n***a*) in their social groups. Thus, members of the UY sample were in a social environment where their orientation toward *n***a*, at least, was relatively positive, and use of the word was socially acceptable among peers. The details of these analyses are available in online materials.

Implicit Attitudes

Implicit attitudes by sample (UY vs. PI) and test (White–Black race attitude IAT vs. Black–*n***er/a* attitude IAT) are shown in Figure 2. A few patterns are notable: First, (mainly White) PI participants showed a stronger implicit preference for White Americans over Black Americans than the UY sample consisting mainly of Black and Hispanic Americans. This difference is in line with previous findings obtained using much larger samples (Morehouse & Banaji, 2024). In fact, implicit race attitudes in the UY sample were, on average, neutral, with the 95-percent confidence interval overlapping with zero. Second, PI participants also demonstrated higher levels of implicit preference for the category Black over *n***er*/*a* than UY participants did. However, implicit preference for Black over the slurs was attenuated but, crucially, not eliminated in the UY sample.

Implicit White–Black race attitudes and implicit Black–*n***er*/*a* attitudes were formally investigated using a linear mixed-effects model with random intercepts for participants, as well as main effects for sample (UY vs. PI), test (White–Black race attitude IAT vs. Black–*n***er*/*a* attitude IAT), and their interaction. The two main effects were significant, $\chi^2(2) = 44.94$, $p < .001$, but the interaction was not, $\chi^2(1) = 0.01$, $p = .918$. The same result was also corroborated

by a Bayesian model selection approach (posterior probability of the main effects model, p = .929).

As expected based on the results reported above, the two samples, $t(299) = 3.58$, $p < .001$, Cohen's $d = 0.41$, $BF_{10} = 33.21$, and the two tests, $t(318) = 2.85$, $p = .005$, Cohen's $d = 0.16$, $BF_{10} = 3.36$, differed from each other significantly. These results have particular significance because, in combination with each other, they suggest that the two samples' differences in implicit negativity toward *n***er/a* are attributable to baseline differences in implicit race attitudes even though they differed in expected ways in their explicitly reported views. The logic behind this analysis is relatively straightforward and is the same one that we used to interpret the results of Study 1: Because both the race IAT and the *n***er/a* IAT rely on White Americans as the contrast category, scores on the race IAT can be used as a baseline to interpret mean-level differences on the *n***er/a* IAT. The fact that the slurs evoke automatic negativity in a sample that explicitly believes the usage to serve a socially positive, affiliative goal, is worth acknowledging as a marker of mean-level dissociation between implicit and explicit cognition.

Secondary Analyses

Next we turned to examining the relationship between implicit and explicit attitudes at the level of individual participants. The overall model with implicit attitudes toward the slurs as the dependent variable and explicit attitudes toward the slurs and toward Black and White Americans as the predictors was statistically significant, $F(4, 251) = 2.62$, $p = .035$. Only explicit attitudes toward *n***a* (but not explicit social group attitudes) had a small but statistically significant effect on implicit attitudes toward *n***er/a*, $\beta = 0.22$, $t(251) = 3.06$, $p = .002$, $BF_{10} = 6.66$, such that more positive explicit attitudes corresponded to more positive implicit attitudes.

Accordingly, the model with explicit attitudes toward *n***a* as the sole predictor had the highest posterior probability in Bayesian model selection framework (p = .698).

In additional analyses reported in detail in online materials, we found that, at the level of individual participants, the race IAT and the *n***er*/*a* IAT were unrelated to each other in both samples, with especially strong evidence for lack of a relationship in the UY sample. Implicit attitudes toward the slurs were also unrelated to self-reported levels of hearing the two words and were weakly correlated with their perceived social acceptability.

Discussion

Study 2 assessed implicit attitudes toward *n***er*/*a* in a sample of theoretical interest: Black and Hispanic underserved youth whose social lives make acceptable, perhaps even demand, the use of the slurs in affiliative contexts. Even in such a sample, implicit attitudes toward the terms remained negative. Moreover, critically, although anti-*n***er*/*a* attitudes were seemingly attenuated toward neutrality in the underserved youth sample (consisting mainly of reappropriators of the terms) relative to the Project Implicit sample (consisting almost entirely of non-reappropriators), this difference can fully be accounted for by baseline differences in implicit race attitudes across the two samples. In other words, although the two groups appear to show different levels of negativity toward the slurs, the difference is entirely explained by reappropriators having weaker implicit race attitudes.

Study 3

Studies 1–2 probed implicit attitudes toward a combination of the two versions of the slur. Although this procedure is defensible for some initial tests, both pre-existing work (D. Smith, 2002; Tesler, 2015) and responses to the explicit items collected in Studies 1–2 attest to the fact that Americans distinguish between the rhotic and non-rhotic forms of the word.

Notably, Black Americans sometimes use *n***a* in expressly positive and affiliative ways within their communities and do so far more frequently than with *n***er*. In fact, the term *n***a* likely evolved to separate it from the version used by the oppressing group.

As such, it is possible that in a direct comparison of the two differences in implicit attitude may emerge. Such tests would be useful to conduct in both populations of volunteer Black and White Americans recruited though channels such as Project Implicit as well as samples consisting mostly of reappropriators. However, samples of reappropriators, such as the underserved youth sample recruited for Study 2, are difficult to obtain for a variety of reasons, most notably that seeking permission to expose individuals to these unique slurs is non-trivial. As such, Study 3 (as well as Studies 4A and 4B below) were conducted in large samples of Project Implicit participants, which made it possible to obtain data from Black Americans who vary in their acceptance of the use of the slurs.

In an initial test reported in the online materials as Supplementary Study 1, we found no difference between Black and White Project Implicit volunteers in their implicit attitudes toward the two words, when each was compared to the socially acceptable contrast term *Black*. However, that study used a between-participant design, which may have made it difficult to detect existing differences due to insufficient power. Here we rely on a more highly powered within-participant design that can take baseline differences between participants into account. Moreover, to be able to appropriately model any effects of race, equal numbers of Black and White American participants were recruited.

In addition to varying the two terms (*n***a* vs. *n***er*) within participants, we manipulated another variable, also within participants: the modality in which the words were presented on the IAT. At the outset of this project, we had expected that the social and emotional power of

the slurs would emerge fully only when participants are exposed to them in spoken form. However, that expectation was yet to be tested empirically. As such, participants in Study 3 completed both a semi-auditory IAT (similar to the ones implemented in Studies 1–2) and a more traditional, purely visual IAT varying the appearance (size, color, and font) of each word on a trial-by-trial basis.

Method

Participants

Equal numbers of Black and White volunteers were recruited from the Project Implicit website. From the pool of participants arriving at Project Implicit, 676 were randomly assigned to the present study. We excluded participants from analyses if they did not complete both IATs, which served as the main dependent measure (n = 23), or had response latencies below 300 ms on 10 percent or more of IAT trials on both IATs, indicating inattention (n = 5). Participants who had usable data from either IAT were retained but data from incomplete IATs or IATs with excessive rapid responding were discarded.

Following participant exclusions, the final dataset consisted of 647 participants. 430 participants identified as female, 203 participants as male, and 7 participants as other genders. Given our recruitment scheme, all participants identified either as Black (n = 326) or as White (n = 321). Mean participant age was 41 years (SD = 15 years). Most participants were politically liberal (n = 316) rather than conservative (n = 132).

Procedure

The procedure for Study 3 was similar to Study 2, with one crucial change. Specifically, participants completed two IATs, each comparing *n***a* and *n***er* directly to each other, one in the auditory and one in the visual modality, in randomized order. After completing the two

IATs, participants responded to three sets of explicit items in randomized order: (*a*) feeling thermometer items measuring explicit attitudes toward Black and White Americans as well as toward the words *n***er* and *n***a*, (*b*) a set of items on the frequency of using and hearing the words, and (*c*) an 8-item scale regarding the acceptability of using the words in their social groups. Finally, participants received feedback on their IAT performance and were debriefed.

The semi-auditory IAT followed the same procedure as IATs in previous studies. Category labels were *N***er* and *N***a,* and category stimuli were voice recordings of the two words. However, instead of using the computer-generated stimuli from Studies 1–2, we recruited Black American volunteers to record themselves saying "*n***er*" and "*n***a*." Six men's and four women's voices were obtained, which kept the total number of stimuli representing each category the same across studies.

On the newly introduced visual IAT, five versions of each word were shown, including "*N***a*," "*n***a*," "*N***A*," "*N***as*," and "*N***AS*," along with "*N***er*," "*n***er*," "*N***R*," "*N***ers*," and "*N***RS*." To create additional variability in the stimuli, one of five visual presentations was randomly selected on each trial, including (*a*) purple, 3em font size, bold; (*b*) green, 2.5em font size, italicized; (*c*) maroon, 4em font size, serif; (*d*) yellow, 3.5em font size, cursive; and (*e*) blue, 2em font size. These variations ensured that participants were not able categorize the stimuli based on visual features alone.

Results

Preliminary Analyses

In a first set of analyses reported in detail in online materials, we ascertained that the two racial groups differed from each other in systematic and expected ways. Indeed, in line with expectations and paralleling the results from Study 2, Black participants were more likely than

White participants to hear and use both *n***er* and *n***a* in their daily lives. Moreover, Black participants expressed less explicit negativity toward the two words, especially *n***a*, than White participants did. Perceived social acceptability of both words (and especially of *n***a*) was also higher among Black than White participants. However, notably, in a deviation from the underserved youth participants from Study 2, Black Project Implicit participants' explicit attitudes toward *n***a* were clearly below the midpoint of the scale.

Implicit Attitudes

Critically, implicit attitudes by race (Black vs. White) and IAT modality (auditory vs. visual) are shown in Figure 3. A few patterns are notable: First, both participant groups and both versions of the IAT exhibited a clear implicit preference for *n***a* over *n***er* on a test providing a direct comparison. Second, the visual test produced more evaluative distinction than the auditory test did. Third, Black participants showed stronger evaluative divergence between the two terms than White participants did.

Implicit attitudes toward *n***a* relative to *n***er* were investigated using a linear mixed-effects model with random intercepts for participants, as well as main effects for participant race (Black vs. White), test modality (auditory vs. visual), and their interaction. The two main effects were significant, $\chi^2(2) = 34.15$, $p < .001$, but the interaction was not, $\chi^2(1) = 0.61$, $p = .433$, posterior probability of the main effects model: $p = .912$.

Overall, these results suggest that (*a*) implicit attitudes toward *n***a* vs. *n***er* were more differentiated among Black relative to White participants; (*b*) unexpectedly, the evaluative difference was stronger on a visual than on an auditory test; and (*c*) the two effects were additive. In follow-up analyses, we investigated the effects of race and modality separately. As expected based on the results reported above, the two racial groups, $t(639) = 3.09$, $p = .002$, Cohen's $d =$

0.24, BF_{10} = 4.93, and the two modalities, $t(615) = 4.99$, $p < .001$, Cohen's $d = 0.40$, BF_{10} = 8717.64, differed significantly from each other.

Secondary Analyses

In additional analyses reported in detail in online materials, we found that the two IATs were moderately related to each other at the participant level ($\beta = 0.39$). Implicit attitudes toward *n***a* relative to *n***er* were unrelated to explicit race attitudes ($\beta = 0.06$ for attitudes toward Black people and $\beta = -0.08$ for attitudes toward White people with the semi-auditory IAT as the dependent measure and $\beta = 0.05$ for attitudes toward Black people and $\beta = -0.02$ for attitudes toward White people with the visual IAT as the dependent measure), but they were weakly positively related to explicit attitudes toward the two words. Specifically, participants showed a stronger implicit preference for *n***a* over *n***er* the more positive their explicit attitudes toward *n***a* ($\beta = 0.23$ for both the semi-auditory and the visual IAT) and the less positive their explicit attitudes toward *n***er* ($\beta = -0.11$ for the semi-auditory IAT and $\beta = -0.13$ for the visual IAT).

Moreover, implicit attitudes were weakly related to self-reported exposure to and use of the words such that participants showed more implicit attitudinal dissociation between the two versions of the term to the extent that they reported hearing *n***a* more frequently and using *n***er* less frequently. Similar to Studies 1–2, older participants showed more evaluative dissociation between the rhotic and non-rhotic variants than younger participants did. No other demographic effects were significant.

Discussion

On a test offering a direct comparison, we found that both Black and White Americans reliably distinguished between the rhotic variant *n***er* and the non-rhotic variant *n***a*,

showing clear implicit preference for the latter over the former. This result was consistent across auditory and visual tests, suggesting that it was not merely a function of the modality in which the words were presented. However, unexpectedly, a stronger effect emerged on the visual than on the semi-auditory IAT. We revisit this difference in Studies 4A–4B.

In addition, we obtained a significant effect of participant race such that Black Americans showed a stronger implicit preference for *n***a* over *n***er* than White Americans did. This result mirrors the results obtained on self-report measures of attitude and familiarity on which Black participants showed more differentiated responses to the two forms of the word relative to White participants. However, despite these similarities in mean levels, like in previous studies, implicit attitudes were at most weakly related to any of the explicit items, including those that conceptually paralleled the IAT, at the level of individual participants.

Studies 4A–4B

Study 3 has provided initial evidence for an implicit preference for *n***a* over *n***er* in a balanced sample of Black and White Americans, attesting to the fact that even subtle differences involving the rhotic and non-rhotic form have taken root in Americans' minds. Notably, this implicit preference was stronger among Black than among White Americans. Moreover, contrary to initial expectations, the effect was more pronounced on a visual than on an auditory test.

However, it is conceivable that the implicit preference for *n***a* over *n***er* emerges only if the two terms are directly contrasted with each other; in comparison to a neutral comparison term (such as *Black*), they may both be overwhelmingly (and equally) negative. As such, a test involving separate contrasts of *n***a* vs. *Black* and *n***er* vs. *Black* is more conservative than the test reported in Study 3. In Studies 4A and 4B, we conduct such a test. In addition, the

results of Studies 1–3 leave the question of the negativity of the two terms, considered separately and in an absolute sense, open. Studies 1–2 have demonstrated that the two terms combined are negative (relative to both *White* and *Black*), and Study 3 has shown that *n***er* is more implicitly negative than *n***a* is. In the final studies, we provide additional evidence on this issue by comparing each term separately to a socially acceptable control with the same referent (*Black*).

We report Studies 4A and 4B together because they followed the same procedure, with the only difference between them being that in Study 4A, participants were recruited without regard to race, whereas Study 4B relied on a racially balanced sample of Black and White Americans. In both studies, participants completed two IATs: one comparing *n***er* to *Black* and the other comparing *n***a* to *Black*, in randomized order.

Modality of the IAT was additionally manipulated between participants such that each participant completed both IATs either in the semi-auditory or in the visual modality. This manipulation served as a test of generalizability and provides further evidence on the auditory–visual difference obtained in Study 3. Given that *Black* is easy to distinguish from both *n***er* and *n***a* auditorily, if the difference observed in Study 3 persists in this study, we can confidently conclude that it is a genuine effect of modality rather than an effect of poor perceptual discriminability.

Method

Participants

Participants for both studies were recruited from the Project Implicit website. In Study 4A, participant race was not restricted, whereas in Study 4B, Black and White participants were recruited in equal numbers to achieve a balanced sample.

From the pool of participants arriving at Project Implicit, 692 were randomly assigned to Study 4A and 807 to Study 4B. We excluded participants from analyses if they did not complete both IATs, which served as the main dependent measure (n = 20 in Study 4A and n = 28 in Study 4B), or had response latencies below 300 ms on 10 percent or more of IAT trials on both IATs, indicating inattention (n = 3 in Study 4A and n = 12 in Study 4B). Participants who had usable data from either IAT were retained but data from incomplete IATs or IATs with excessive rapid responding were discarded.

Following participant exclusions, the final dataset in Study 4A consisted of 669 participants. 452 participants identified as female, 209 participants as male, and 5 participants as other genders. Most participants identified as White (n = 363), followed by self-identification as Black (n = 92), Hispanic (n = 90), and Asian (n = 57). Mean participant age was 38 years (SD = 14 years). Most participants were politically liberal (n = 320) rather than conservative (n = 146).

The final dataset in Study 4B consisted of 767 participants. 524 participants identified as female, 220 participants as male, and 21 participants as other genders. Given our recruitment scheme, all participants identified either as Black (n = 390) or as White (n = 377). Mean participant age was 36 years (SD = 16 years). Most participants were politically liberal (n = 377) rather than conservative (n = 125).

Procedure. The procedure for both studies was the same, with the only difference consisting in the racial composition of the sample. All participants completed two IATs, one of them comparing *n***a* to Black and the other one comparing *n***er* to Black, in randomized order. In addition, the modality of the two tests (semi-auditory vs. visual) was manipulated between participants such that each participant completed both IATs in the same modality. The procedure for both types of IAT was the same as in Study 3. After completing both IATs, participants

responded to the same explicit items as participants in Study 3 had. Finally, participants received feedback on their IAT performance and were debriefed.

Results

Preliminary Analyses

In a first set of analyses reported in detail in online materials, we ascertained that the two racial groups differed from each other in systematic and expected ways. All results involving explicit attitudes, use of and exposure to the two words, and social acceptability reported in Study 3 were replicated in both Studies 4A and 4B and are therefore not discussed further.

Implicit Attitudes

Critically, implicit attitudes by contrast (Black–*n***a* vs. Black–*n***er*), IAT modality (auditory vs. visual), and, for Study 4B, race (Black vs. White) are shown in Figures 4 and 5. A few patterns are notable: First, participants in all conditions of both studies showed a preference for Black over *n***a* and *n***er*. Second, implicit negativity toward *n***er* was stronger than implicit negativity toward *n***a*, replicating the effect obtained in Study 3 using a more conservative test. Third, the effect was stronger on a visual than on an auditory test, again replicating the result obtained in Study 3. Finally, Black participants showed stronger evaluative divergence between the two terms than White participants did, also replicating the result from Study 3.

In Study 4A, implicit attitudes toward the slurs relative to the contrast category Black were investigated using a linear mixed-effects model with random intercepts for participants, as well as main effects for comparison (*n***er* vs. Black or *n***a* vs. Black), test modality (auditory vs. visual), and their interaction. The two main effects were significant, $\chi^2(2) = 25.53$, $p < .001$, but the interaction was not, $\chi^2(1) = 0.009$, $p = .927$, posterior probability of the main effects model: $p = .651$. Follow-up analyses indicated that the effect was stronger for *n***er* than for

*n***a* (relative to Black), $t(629) = 3.71$, $p < .001$, Cohen's $d = 0.30$, $BF_{10} = 38.60$, and in the visual than in the auditory modality, $t(661) = 3.49$, $p < .001$, Cohen's $d = 0.27$, $BF_{10} = 17.60$.

In Study 4B, we fit the same model but added participant race (Black vs. White) as another independent variable. Similar to Study 4A, the main effects of comparison and modality were significant, with an additional main effect of race, $\chi^2(3) = 37.97$, $p < .001$, posterior probability of the main effects model: $p = .688$. Follow-up analyses indicated that the effect was stronger for *n***er* than for *n***a*, $t(725) = 3.82$, $p < .001$, Cohen's $d = 0.28$, $BF_{10} = 55.23$, in the visual than in the auditory modality, $t(756) = 3.71$, $p < .001$, Cohen's $d = 0.27$, $BF_{10} = 37.30$, and among Black than among White participants, $t(755) = 3.31$, $p < .001$, Cohen's $d = 0.24$, $BF_{10} = 9.34$. Given that no interactions were significant, the three effects (slur variant, modality of presentation, and participant race) were additive.

Secondary Analyses

In additional analyses reported in detail in online materials, we found that the two IATs were moderately related to each other at the level of individual participants ($\beta = 0.37$), replicating the result obtained in Study 3. Also replicating the results from Study 3, implicit attitudes were not predicted by self-reported usage or exposure, by explicit attitudes toward the two terms or toward Black and White Americans, or perceived acceptability of using each word in one's social circles. In Study 4B, explicit attitudes toward Black people and toward *n***a* produced small and inconsistent effects; even these weak effects did not emerge in Study 4A. Similar to Studies 1–3, younger participants showed less implicit negativity toward the slurs than older participants did. No other demographic effects were significant.

Discussion

Studies 4A–4B replicated the main results of Study 3 using a more conservative test: American participants' implicit attitudes consistently distinguished between the words *n***a* and *n***er*, with a clear implicit preference for the former over the latter, this time relative to a socially acceptable control term with the same referent. This result provides evidence for the robustness of the main finding of Study 3. In addition, the modality effect observed there was also replicated, suggesting that the stronger evaluative dissociation between the two versions of the slur on the visual compared with the auditory test could not have been a result of lack of discriminability on the former given that the present study contrasted *Black* and *n***er* as well as *Black* and *n***a*, with both pairs of stimuli easily distinguishable from each other.

Also similar to Study 3, Black participants exhibited stronger implicit negativity toward both terms than White participants did. Crucially, the results of Study 4B are difficult to reconcile with claims of reappropriation among Black Americans given that Black participants exhibited more rather than less implicit negativity toward *n***a* relative to White participants. This result is conceptually in line with the results of Study 2 where, once the appropriate controls were taken into account, even underserved Black and Hispanic youth (conscious reappropriators) showed the same amount implicit negativity toward the two terms combined as White participants did.

General Discussion

Debates about the meaning and legitimacy of the slurs *n***er* and *n***a* are alive in American society today. And yet, virtually nothing is known about how these terms are represented in the minds of Black and White Americans given that no experimental investigation of the automatic cognitive–affective status of the terms has ever been conducted. In the present studies, we used a measure of implicit cognition to create a window into how the valence of

these terms operates automatically. When the two terms were tested in combination relative to socially acceptable controls, such as *Black* or *White*, measures of implicit attitudes revealed strong negativity. Notably, such negativity emerged even in a sample of young Black and Hispanic students who are conscious reappropriators of the terms, especially of *n***a*. Critically, these students' weaker implicit negativity toward *n***er* and *n***a* was entirely attributable to having weaker implicit race attitudes than a sample of (mostly White) volunteer participants. In other words, the slurs remain squarely negative in meaning today.

This is not to say, however, that implicit attitudes did not distinguish between the two forms of the term. *N***er* was found to be associated with stronger implicit negativity than *n***a* was, both using a direct comparison and when each was separately contrasted with *Black*. Notably, Black Americans distinguished more between the two terms than White Americans did and showed especially high levels of negativity toward *n***a*, although on self-report measures they expressed more positivity, higher frequency of exposure, and higher social acceptability. This result represents a direct counterpoint to the previous finding suggesting that the word remains negative in meaning. Here we see adaptation such that a new version (*n***a*) created for the purpose of reappropriation has had some impact: Both White and Black Americans — but especially Black Americans — are sensitive to a distinction involving no more than the final syllable of a word. As time passes, it will be of interest to study whether *n***a* has fully transformed to be positive in meaning in absolute terms, which is currently not the case.

Unexpectedly, both on a direct comparison and comparisons involving a socially acceptable contrast term with the same referent, both Black and White Americans exhibited stronger implicit negativity toward *n***er* and *n***a* when presented visually rather than auditorily. This finding contradicted our initial conjecture according to which negative implicit attitudes should

be especially pronounced when hearing (rather than seeing) the slurs. This result, emerging from Studies 3–4B, is a good example of naïve theories being incorrect. At present, we do not have a compelling explanation for the power of the visual over the auditory version of the slurs, but the strength of the result and its robustness across studies invites more research to reach a better understanding.

Here we can offer only some initial speculations as to why the modality effect may have emerged and leave more systematic tests for future work. Although, at least among White Americans, encountering the terms is rare, participants are probably more likely to hear rather than read them in their daily lives. As such, the visual versions used on the IAT may have produced more of a surprise, and consequently, a stronger evaluative effect, than the auditory versions did. Notably, on the auditory version, the words were pronounced either by computer-generated voices or voices of Black Americans. Had the slurs been read aloud by White Americans (for whom use of the words is unquestionably taboo), the modality effect may have been attenuated or even reversed. Moreover, probing implicit attitudes evoked by other slurs in the process of reappropriation (such as *q*eer*, *f*ggot*, or *b*tch*) could inform about whether the modality effect is specific to *n***er/a*, or whether it emerges more broadly.

Theoretical and Practical Implications

The present findings have implications for the nature and evolution of slurs, the process of reappropriation, as well as the nature of implicit attitudes more generally. We discuss each of these implications below.

First, across all five studies, implicit attitudes toward the terms were found to be at most weakly related to self-report measures of (*a*) attitudes toward them, (*b*) attitudes toward Black and White Americans, (*c*) frequency of exposure, (*d*) frequency of use, or (*e*) social acceptability.

This finding highlights the importance of studying the automatic aspects of evaluation related to slur words that are in the process of being reappropriated by members of the affected communities. Indeed, on the basis of self-report measures alone we would have concluded that the word *n***a* has become more positive than not among Black Americans in general, and unequivocally positive among the underserved youth sample consisting mostly of conscious reappropriators. However, as far as implicit attitudes are concerned, Black participants were equally if not more negative toward both versions of the slur than White Americans were.

At the same time, these findings should not be understood to indicate that implicit attitudes toward the terms are set in stone. Indeed, an emerging line of work has provided evidence for the remarkable malleability of implicit attitudes not only in the lab (Cone et al., 2017; Kurdi & Dunham, 2020) but even over longer time periods in American culture at large (Charlesworth & Banaji, 2019; for a recent review, see Kurdi & Charlesworth, 2023). In fact, implicit anti-Black attitudes declined at rates faster between 2017 and 2020 than originally predicted based on data from the preceding decade (Charlesworth & Banaji, 2022a), presumably in response to increased awareness of structural racism and bias. As such, it is conceivable that reappropriation of *n***a* might occur in the underserved youth sample with usage over longer time periods and continued infusion of positive meaning in a society that is also showing some reduction in implicit anti-Black attitudes.

However, it is notable that the shift toward neutrality in implicit race attitudes has affected virtually all segments of the U.S. population, encompassing Americans of all ages, races, genders, and educational levels (Charlesworth & Banaji, 2021). This result, in turn, suggests that change may have been possible due to macro-level factors present in all (or at least most) segments of society, such as the Black Lives Matter movement, which resulted in higher levels of

interest in and awareness of historical and present-day anti-Black racism in the United States (Barrie, 2020; Reny & Newman, 2021). By contrast, reductions in implicit negativity toward slur words, such as *n***a*, among members of certain racial and ethnic communities may occur more slowly or potentially not at all. The reason for this is that macro-level factors, such as social sanctions against using the terms, are likely to counteract the effects of attempts at conscious re-appropriation at the micro-level, including use in affiliative contexts of family and friendship.

Second, the robust relationship between participant age and implicit attitudes raises another set of questions about the lifespan of reappropriation. Specifically, the fact that participant age was consistently and positively correlated with implicit negativity toward both *n***a* and *n***er* across all studies, along with the result that the youngest sample showed the least negativity toward the words overall, implies that we might be capturing a snapshot of the reappropriation process unfolding. If we could go back in time and measure implicit attitudes toward *n***er* and *n***a* repeatedly, starting in the late 1980s with the introduction of the gangsta rap group, N.W.A. (N***az Wit Attitudes), would we find that attitudes toward the words have gotten more positive over time?

We do not have these data at present, but computational methods, such as word embedding algorithms (Caliskan & Lewis, 2020; Charlesworth et al., 2022; Charlesworth & Banaji, 2022b), may make such longitudinal analyses of archival text data possible. In fact, change in attitudes toward the terms could be reflective of an aging effect (i.e., most positive attitudes when individuals are in their youth and least positive in older age) rather than a cohort effect (i.e., a change characterizing a population born at a particular time; Blanchard et al., 1977). To disentangle these possibilities, we hope that future research will continue to measure these attitudes and track the evolution of valence over time.

Third, the mechanisms mediating the connection between usage and implicit attitudes remain opaque, especially in the realm of individuals using *n***a* in self-referent or affiliative ways to challenge racial stigma or to reclaim a sense of power and autonomy over racial issues (Galinsky et al., 2013; Wang et al., 2017). Understanding this connection becomes even more challenging when one considers the fact that while Black Americans can say *n***a* affectionately, they also use it derogatorily. For example, when *n***a* is used as a term of endearment, it often includes self-reference, such as in "*my n***a*." However, referring to another simply as "*a n***a*" is usually derogatory and invokes stereotypes about criminal behavior, infidelity, and laziness.

Furthermore, the context in which the word is used or heard likely impacts the valence that it evokes. For example, implicit attitudes may be positive when the term is used in a familial environment, as opposed to in the workforce (Rosette et al., 2013). Given that considerable evidence for the context-dependence of implicit attitudes (Gawronski et al., 2018), and specifically implicit race attitudes (Wittenbrink et al., 2001), exists, future work should investigate how processes of reappropriation and self-slur labeling contribute to changes in implicit attitudes toward the words and whether such effects show variability depending on the context in which implicit attitudes are elicited.

Fourth, the present findings invite inquiry into whether *n***er* and *n***a* are truly unique words regarding implicit attitudes and trajectory of reappropriation, or whether other slurs that have been reappropriated or are in the process of reappropriation (e.g., *b*tch*, *q*eer*, and *f*ggot*) show a similar dissociation between explicit and implicit attitudes. For example, in one study researchers found that subliminally priming straight participants with *f*ggot*, compared to *gay*, led to interference with their ability to categorize positive words (Carnaghi &

Maas, 2007). In contrast, gay participants were unaffected by the primes. Relatedly, subliminal and supraliminal exposure to homophobic slurs resulted in increased dehumanization of, and distancing from, gay men among straight participants (Fasoli et al., 2016). Future work should directly compare *n***er* and *n***a* with other reappropriated words, expanding the scope of our understanding of linguistic shifts over time as well as the unique representation of slurs and their reappropriation.

Fifth, the present studies also speak to more general accounts of how implicit attitudes are acquired and maintained in the human mind. Specifically, under some associative theories (McConnell & Rydell, 2014; Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000), implicit attitudes are thought to reflect the piecemeal accumulation in a slow-learning mental system of evaluative associations experienced with an attitude object over longer periods of time. The present results seem difficult to reconcile with theories of this kind: The term *n***er* is rarely, if ever, heard or read in present-day American society and yet it continues to evoke highly negative valence in most White and Black Americans. One possibility, broadly in line with propositional accounts of implicit cognition (De Houwer, 2014), is that despite the disuse of *n***er*, there is sufficient conceptual knowledge about the history and meaning of the word to keep negative valence alive, even at an implicit level. Alternatively, or in addition, strong social sanctions against using these words, which are present in most segments of American society today, may be sufficient to attach extremely negative valence to them, even in the absence of regular reinforcement (Sarin et al., 2021).

However, if — as we argue above — both implicit and explicit attitudes are sensitive to propositional content going well beyond simple stimulus–stimulus and stimulus–response associations, why did we observe at most weak relationships between the IAT and even parallel self-

report measures in the present studies? We believe that at least two possibilities are worthy of consideration. First, it is possible that although both sets of measures reflect the same evaluative representations, self-report measures are also sensitive to other, not directly attitudinal processes, such as awareness of societal norms or self-presentational concerns (Fazio, 2007). Second, alternatively or additionally, it is conceivable that although both the IAT and self-report measures index propositional content, such propositional content is retrieved only partially on the former and more fully on the latter (Van Dessel et al., 2019). Under both accounts, implicit–explicit dissociations can occur without positing different learning mechanisms giving rise to each. Providing direct evidence for these (or other) possibilities may be a fruitful avenue for future work.

Limitations and Directions for Follow-Up Work

Above, we have already mentioned several possibilities for potential follow-up work. We conclude by expanding on each of these possibilities and highlighting some further ones.

First, the present project provides a snapshot of (implicit) attitudes toward *n***er* and *n***a* at a particular point in cultural time. Follow-up work could track how relevant attitudes among both Black and White Americans may shift in response to cultural and historical trends, with the present results serving as a point of reference. In addition, computational methods, such as word embeddings, could be fruitfully used to uncover not only future developments but to also probe how attitudes may already have changed over time (see, e.g., Charlesworth et al., 2022).

Second, we have repeatedly touched upon the importance of context effects, which the present work was not directly designed to test, although it did explore variability in stimulus materials (e.g., auditory stimulus presentation in Studies 1–2 vs. visual stimulus presentation in Studies 3–4B; computer-generated voices using standard American English in Studies 1–2 vs. human voices of Black American speakers in Studies 3–4B; rhotic vs. non-rhotic versions of the

term in Studies 3–4B; and a number of different contrast terms across Studies 1–4B). Specifically, future work might more systematically probe whether attitudes toward the slurs differ depending on the (*a*) physical and social setting of use, (*b*) the identity of the speaker and their interaction partner, and (*c*) a number of other factors. Natural language processing algorithms that allow for sentence-level analyses (such as BERT; Devlin et al., 2018) may be able to shed some light on whether and to what degree the valence and other psychological phenomena associated with the terms varies across social contexts.

Third, although the present studies relied both on diverse samples of American adults and a targeted sample of young Black reappropriators, the samples were not representative. Specifically, given that Project Implicit relies on volunteers who are willing to spend their time participating in studies on social group-based biases, the relevant samples may not be able to capture the full range of variability on individual differences that may contribute to implicit attitudes toward the slurs. For example, White Americans endorsing extreme right views may be comfortable using *n***er* and *n***a* and to report such views to the experimenter. As such, future work may benefit from attempting to replicate the present findings in more representative samples.

Fourth, the present work is limited by the use of a single dependent measure. Although the IAT was selected due to its strong psychometric properties and its ease of administration, it is subject to a number of constraints. Notably, the IAT measures evaluative content in relative ways, which makes it difficult to draw inferences about stimulus evaluations in the absolute sense. Related to the previous point, it is possible that (certain subgroups of) participants may have felt ambivalent toward the slurs; however, to measure such ambivalence, future work will have to rely on alternative sets of self-report (Ng et al., 2022) and indirect (Zayas et al., 2022) measures.

***

In an interview with Randall Kennedy, upon the publication of his book *Nigger: The Strange Career of a Troublesome Word*, author Daniel Smith states that "[…] nigger is far from static in meaning. It can connote vitriol, yes, but it can also connote camaraderie. It can be said angrily, but it can also be said with irony. It can be thrown like a grenade, but it can also be picked up and thrown back. Just as there are devastatingly bad uses of nigger, there are, Kennedy believes, 'good uses'—uses that can promote the cause of justice (Mark Twain's bitterly facetious "Only a Nigger") or that can help 'yank *nigger* away from white supremacists' (the comedy of Richard Pryor and Chris Rock). And in Kennedy's opinion, we are moving in the right direction." (D. Smith, 2002)

It is possible that the movement is in the right direction. However, although *n***a* appears to be on a path to transformation, as of now, the automatic meaning elicited by both terms remains one of clear negative valence in both Black and White Americans' minds. Both of these results are important. The latter result suggests that history casts a long shadow to keep even disused words active and alive in their original (negative) meaning. The former result, in turn, shows the capacity of human minds to adapt to new meaning that is culturally created. In this case, even a minor phonological shift from one version (*n***er*) to another (*n***a*) is associated with a sizeable difference in automatically evoked connotations, with the non-rhotic version *n***a* subject to considerably higher levels of implicit positivity. As such, *n***a* appears to have begun its journey toward reappropriation.

References

Banaji, M. R. & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, *7*(3), 136–141. https://doi.org/10.1111/j.1467-9280.1996.tb00346.x

Bar-Anan, Y. & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*(3), 668–688. https://doi.org/10.3758/s13428-013-0410-6

Baron, A. S. & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science*, *17*(1), 53–58. https://doi.org/10.1111/j.1467-9280.2005.01664.x

Barrie, C. (2020). Searching racism after George Floyd. *Socius*, *6*. https://doi.org/10.1177/2378023120971507

Bauer-Wolf, J. (2018, January 23). Kicked out for racism. *Inside Higher Ed*. https://www.insidehighered.com/news/2018/01/23/university-alabama-may-have-violated-first-amendment-kicking-out-racist-student

Bauer-Wolf, J. (2019, April 9). Anger over n-word at American U. *Inside Higher Ed*. https://www.insidehighered.com/news/2019/04/09/video-american-university-student-using-racial-slur-goes-viral

Blanchard, R. D., Bunker, J. B. & Wachs, M. (1977). Distinguishing aging, period and cohort effects in longitudinal studies of elderly populations. *Socio-Economic Planning Sciences*, *11*(3), 137–146. https://doi.org/10.1016/0038-0121(77)90032-5

Caliskan, A. & Lewis, M. (2020). *Social biases in word embeddings and their relation to human cognition*. PsyArXiv. https://psyarxiv.com/d84kg/

Carnaghi, A. & Maas, A. (2007). In-group and out-group perspectives in the use of derogatory group labels. *Journal of Language and Social Psychology*, *26*(2), 142–156. https://doi.org/10.1177/0261927x07300077

Charlesworth, T. E. S. & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science*, *30*(2), 174–192. https://doi.org/10.1177/0956797618813087

Charlesworth, T. E. S. & Banaji, M. R. (2021). Patterns of implicit and explicit attitudes II. Long-term change and stability, regardless of group membership. *American Psychologist*, *76*(6), 851–869. https://doi.org/10.1037/amp0000810

Charlesworth, T. E. S. & Banaji, M. R. (2022a). Patterns of implicit and explicit attitudes: IV. 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 Inc.

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), e2121798119. https://doi.org/10.1073/pnas.21217981191

Coates, T.-N. (2013, November 23). In defense of a loaded word. *The New York Times*. https://www.nytimes.com/2013/11/24/opinion/sunday/coates-in-defense-of-a-loaded-word.html

Cone, J., Mann, T. C. & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. *Advances in Experimental Social Psychology* (Vol. 56, pp. 131–199). Elsevier Inc. https://doi.org/10.1016/bs.aesp.2017.03.001

Cvencek, D., Meltzoff, A. N. & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development*, *82*(3), 766–779. https://doi.org/10.1111/j.1467-8624.2010.01529.x

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342–353. https://doi.org/10.1111/spc3.12111

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

Devitt, M., & Sterelny, K. (1999). *Language and Reality: An Introduction to the Philosophy of Language* (2nd ed.). MIT Press.

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

Fasoli, F., Paladino, M. P., Carnaghi, A., Jetten, J., Bastian, B. & Bain, P. G. (2016). Not "just words": Exposure to homophobic epithets leads to dehumanizing and physical distancing from gay men. *European Journal of Social Psychology*, *46*(2), 237–248. https://doi.org/10.1002/ejsp.2148

Fazio, R. H. (2007). Attitudes as object–evaluation associations of varying strength. *Social Cognition*, *25*(5), 603–637. https://doi.org/10.1521/soco.2007.25.5.603

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C. & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Galinsky, A. D., Wang, C. S., Whitson, J. A., Anicich, E. M., Hugenberg, K. & Bodenhausen, G. V. (2013). The reappropriation of stigmatizing labels. *Psychological Science*, *24*(10), 2020–2029. https://doi.org/10.1177/0956797613482943

Garcia, J. L. A. (2003). Nigger: The strange career of a troublesome word. *Society*, *40*(5), 93–96. https://doi.org/10.1007/bf03008268

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B. & Hu, X. (2018). Contextualized attitude change. *Advances in Experimental Social Psychology* (Vol. 57, pp. 1–52). Elsevier. https://doi.org/10.1016/bs.aesp.2017.06.001

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., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., Hahn, A., Hehman, E., Hofmann, W., Hughes, S., Hussey, I., Jordan, C., Kirby, T. A., Lai, C. K., Lang, J. W. B., Lindgren, K. P., Maison, D., Ostafin, B. D., Rae, J. R., … Wiers, R. W. (2022). Best research practices for using the Implicit Association Test. *Behavior Research Methods*, *54*(3), 1161–1180. https://doi.org/10.3758/s13428-021-01624-3

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

Greenwald, A. G., Nosek, B. A. & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037//0022-3514.85.2.197

Haaf, J. M. & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*(4), 779–798. https://doi.org/10.1037/met0000156

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., Foels, R. & Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO7 scale. *Journal of Personality and Social Psychology*, *109*(6), 1003–1028. https://doi.org/10.1037/pspi0000033

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596x(91)90025-f

Kennedy, R. L. (1999). Who can say "nigger"? And other considerations. *The Journal of Blacks in Higher Education*, *26*, 86. https://doi.org/10.2307/2999172

Kirkland, P. & Silverman, G. (2018, December 28). The N-word project. *The Washington Post*. https://www.washingtonpost.com/graphics/lifestyle/the-n-word/

Kurdi, B. & Banaji, M. R. (2017). Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *Journal of Experimental Psychology: General*, *146*(2), 194–213. https://doi.org/10.1037/xge0000239

Kurdi, B. & Charlesworth, T. E. S. (2023). A 3D framework of implicit attitude change. *Trends in Cognitive Sciences*, *27*(8), 745–758. https://doi.org/10.1016/j.tics.2023.05.009

Kurdi, B. & Dunham, Y. (2020). Propositional accounts of implicit evaluation: Taking stock and looking ahead. *Social Cognition*, *38*(Supplement), s42–s67. https://doi.org/10.1521/soco.2020.38.supp.s42

Kurdi, B., Morehouse, K. N. & Dunham, Y. (2022). How do explicit and implicit evaluations shift? A preregistered meta-analysis of the effects of co-occurrence and relational information. *Journal of Personality and Social Psychology*, *124*(6), 1174–1202. https://doi.org/10.1037/pspa0000329

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). Academic Press.

McConnell, A. R. & Rydell, R. J. (2014). The systems of evaluation model: A dual-systems approach to attitudes. In J. W. Sherman, B. Gawronski & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 204–217). Guilford Press.

McWhorter, J. (2010, August 17). The Root: Let's make a deal on the n-word. *National Public Radio*. https://www.npr.org/2010/08/17/129250175/the-root-lets-make-a-deal-on-the-n-word

Meyer, D. E. & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. https://doi.org/10.1037/h0031564

Montgomery, B. & Cartwright, L. (2021, February 5). NY Times star reporter Donald McNeil exits after Daily Beast exposé. *Daily Beast*. https://www.thedailybeast.com/ny-times-star-reporter-fired-after-daily-beast-expose

Morehouse, K. N. & Banaji, M. R. (2023). The science of implicit race bias: Evidence from the Implicit Association Test. *Daedalus*, *153*(1), 21–50. https://doi.org/10.1162/daed_a_02047

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*(5), 648–654. https://doi.org/10.3758/bf03213230

Ng, W. J. R., See, Y. H. M. & Wallace, L. E. (2022). When objective ambivalence predicts subjective ambivalence: An affect–cognition matching perspective. *Personality and Social Psychology Bulletin*, *49*(10), 1495–1510. https://doi.org/10.1177/01461672221102015

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G. & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36–88. https://doi.org/10.1080/10463280701489053

Parise, C. V. & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: A study using the Implicit Association Test. *Experimental Brain Research*, *220*(3–4), 319–333. https://doi.org/10.1007/s00221-012-3140-6

Rahman, J. (2012). The n word: Its history and use in the African American community. *Journal of English Linguistics*, *40*(2), 137–171. https://doi.org/10.1177/0075424211414807

Ratliff, K. A., Lofaro, N., Howell, J. L., Conway, M. A., Lai, C. K., O'Shea, B., Smith, C. T., Jiang, C., Redford, L., Pogge, G., Umansky, E., Vitiello, C. A. & Zitelny, H. (2020). *Documenting bias from 2007–2015: Pervasiveness and correlates of implicit attitudes and stereotypes II*. PsyArXiv. https://osf.io/jeyc7

Reny, T. T. & Newman, B. J. (2021). The opinion-mobilizing effect of social protest against police violence: Evidence from the 2020 George Floyd protests. *American Political Science Review*, *115*(4), 1499–1507. https://doi.org/10.1017/s0003055421000460

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D. & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*(6), 1295–1309. https://doi.org/10.1177/1745691620927709

Rosette, A. S., Carton, A. M., Bowes-Sperry, L. & Hewlin, P. F. (2013). Why do racial slurs remain prevalent in the workplace? Integrating theory on intergroup behavior. *Organization Science*, *24*(5), 1402–1421. https://doi.org/10.1287/orsc.1120.0809

Rouder, J. N., Haaf, J. M. & Aust, F. (2018). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, *85*(1), 41–56. https://doi.org/10.1080/03637751.2017.1394581

Rydell, R. J. & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*(6), 995–1008. https://doi.org/10.1037/0022-3514.91.6.995

Sarin, A., Ho, M. K., Martin, J. W. & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, *208*, 104544. https://doi.org/10.1016/j.cognition.2020.104544

Sellers, R. M., Smith, M. A., Shelton, J. N., Rowley, S. A. J. & Chavous, T. M. (1998). Multidimensional model of racial identity: A reconceptualization of African American racial identity. *Personality and Social Psychology Review*, *2*(1), 18–39. https://doi.org/10.1207/s15327957pspr0201_2

Skinner, B. F. (1958). Reinforcement today. *American Psychologist*, *13*(3), 94–99. https://doi.org/10.1037/h0049039

Smith, D. (2002). That word. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2002/01/that-word/303059/

Smith, E. R. & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*(2), 108–131. https://doi.org/10.1207/s15327957pspr0402_01

Tesler, M. (2015, June 25). Using the n-word is more common than you (or President Obama) may think. *The Washington Post*. https://www.washingtonpost.com/news/monkey-cage/wp/2015/06/25/using-the-n-word-is-more-common-than-you-or-president-obama-may-think/

Van Dessel, P., Gawronski, B. & De Houwer, J. (2019). Does explaining social behavior require multiple memory systems? *Trends in Cognitive Sciences*, *23*(5), 368–369. https://doi.org/10.1016/j.tics.2019.02.001

Wang, C. S., Whitson, J. A., Anicich, E. M., Kray, L. J. & Galinsky, A. D. (2017). Challenge your stigma. *Current Directions in Psychological Science*, *26*(1), 75–80. https://doi.org/10.1177/0963721416676578

Wittenbrink, B., Judd, C. M. & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology*, *81*(5), 815–827. https://doi.org/10.1037//0022-3514.81.5.815

Zayas, V., Wang, A. M. & McCalla, J. D. (2022). Me as good and me as bad: Priming the self triggers positive and negative implicit evaluations. *Journal of Personality and Social Psychology*, *122*(1), 106–134. https://doi.org/10.1037/pspp0000332

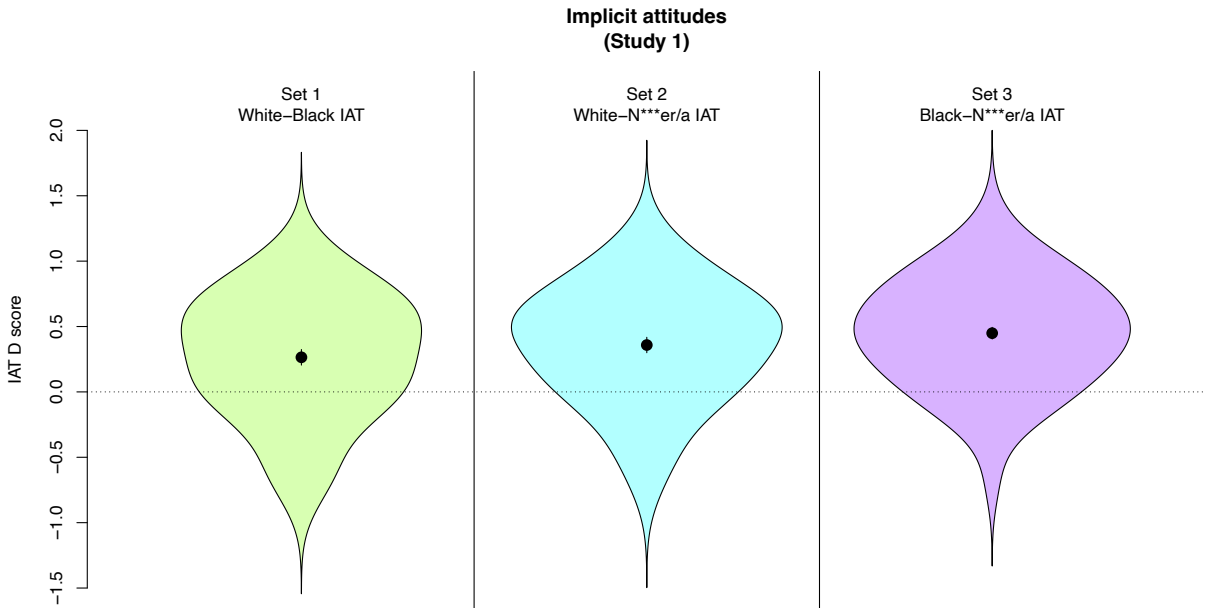

Figure 1. Implicit attitudes toward the categories White vs. Black (set 1), White vs. *n***er/a* (set 2), and Black vs. *n***er/a* (set 3; Study 1). For the specific description of category labels as well as mean levels of responding on each IAT separately, see online materials. Solid black dots correspond to means and error bars correspond to 95-percent confidence intervals. The dashed horizontal line represents neutrality. Positive scores indicate an implicit preference for the first over the second category in each hyphenated label.

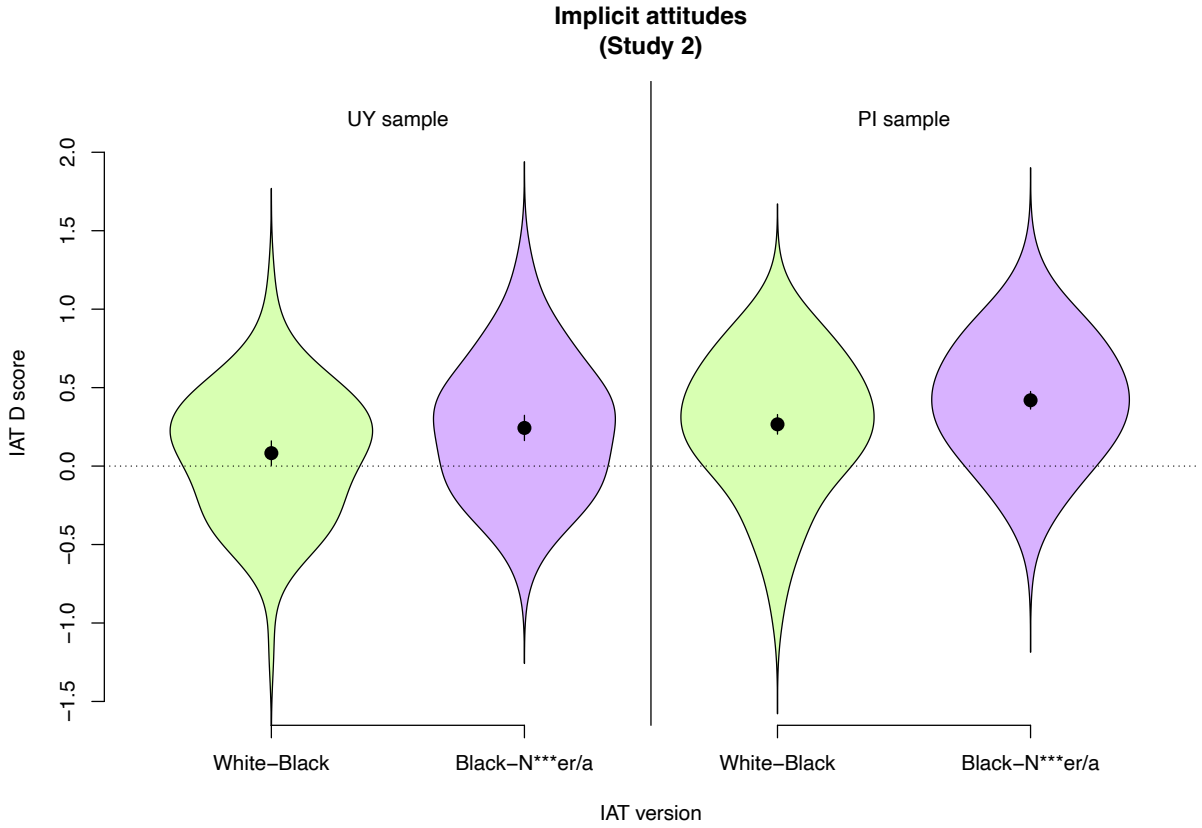

Figure 2. Implicit attitudes toward the categories White vs. Black and Black vs. *n***er/a* by participant sample (Study 2). Solid black dots correspond to means and error bars correspond to 95-percent confidence intervals. The dashed horizontal line represents neutrality. Positive scores indicate an implicit preference for the first over the second category. UY = underserved youth; PI = Project Implicit.

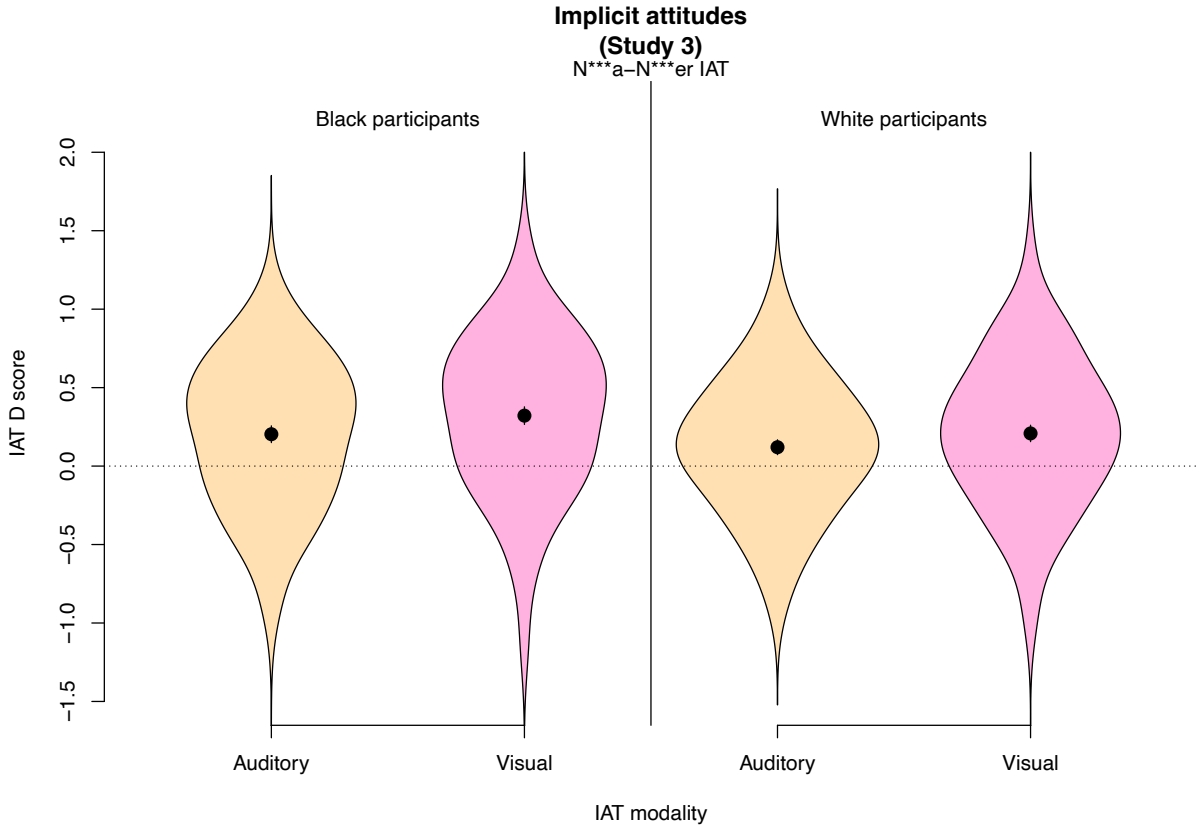

Figure 3. Implicit attitudes toward the categories *n***a* vs. *n***er* by participant sample and modality of test (Study 3). Solid black dots correspond to means and error bars correspond to 95-percent confidence intervals. The dashed horizontal line represents neutrality. Positive scores indicate an implicit preference for the first over the second category.

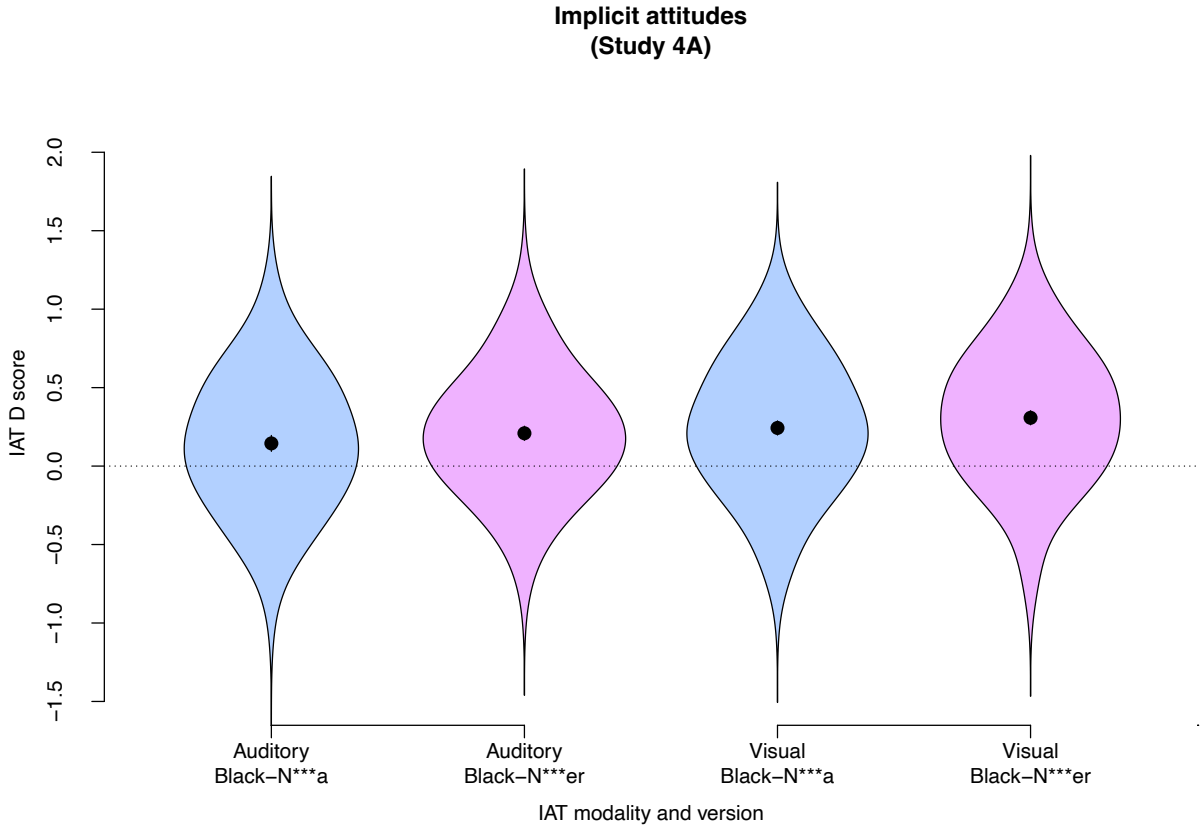

Figure 4. Implicit attitudes toward the categories Black vs. *n***a* and Black vs. *n***er* by modality of test (Study 4A). Solid black dots correspond to means and error bars correspond to 95-percent confidence intervals. The dashed horizontal line represents neutrality. Positive scores indicate an implicit preference for the first over the second category.

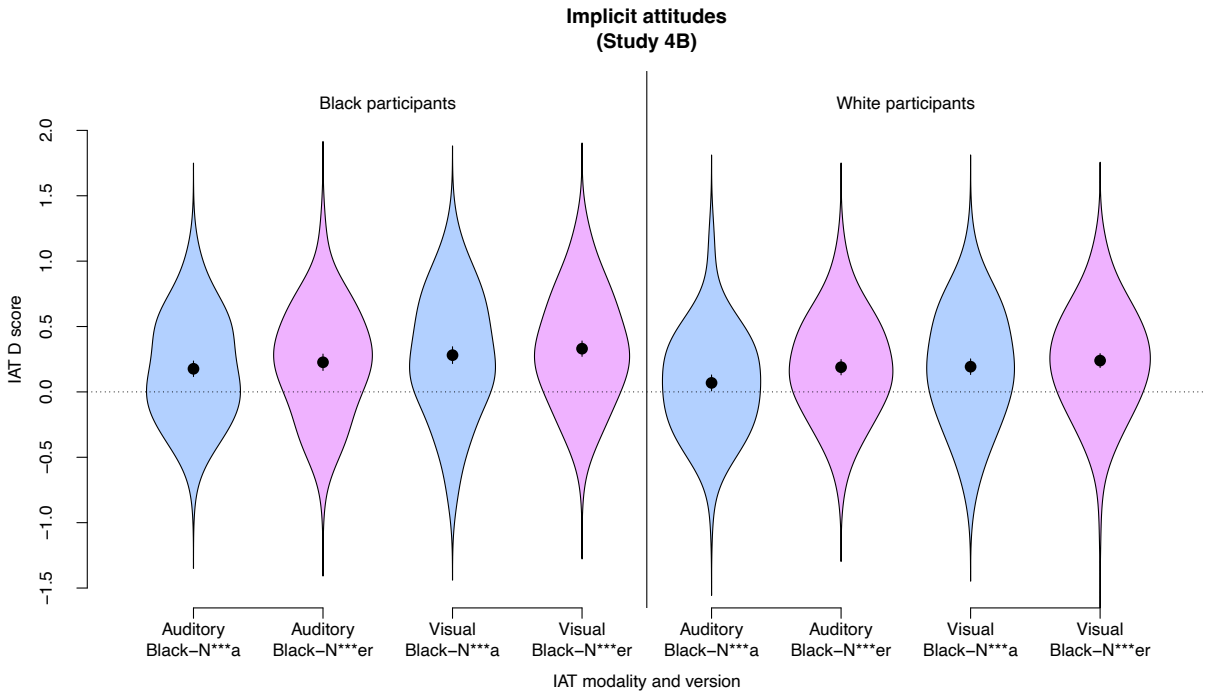

Figure 5. Implicit attitudes toward the categories Black vs. *n***a* and Black vs. *n***er* by sample and modality of test (Study 4B). Solid black dots correspond to means and error bars correspond to 95-percent confidence intervals. The dashed horizontal line represents neutrality. Positive scores indicate an implicit preference for the first over the second category.

Study	Sample	Final *N*	Number of IATs	IAT comparisons	IAT modality
Study 1	Project Implicit, all races	817	1 per participant	White–Black (2 versions), White–*N***er/a* (3 versions), Black–*N***er/a* (3 versions)	Auditory
Study 2	Underserved youth (UY), Project Implicit (PI)	326	2 per participant	White–Black and Black–*N***er/a*	Auditory
Study 3	Project Implicit, Black and White samples	647	2 per participant	*N***a*–*N***er*	Auditory and visual
Study 4A	Project Implicit, all races	669	2 per participant	Black–*N***a* and Black–*N***er*	Auditory or visual
Study 4B	Project Implicit, Black and White samples	767	2 per participant	Black–*N***a* and Black–*N***er*	Auditory or visual

Table 1. Overview of the samples and designs of Studies 1–4. IAT = Implicit Association Test.