Journal of Experimental Social Psychology 123 (2026) 104842

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Registered Report Stage 2: Full Article

The link between social categorization and spontaneous social evaluations: A matter of the evaluative implications of the situation?☆

Manuel Becker [a,b,*], Sarah Teige-Mocigemba [b], Jeffrey W. Sherman [c], Karl Christoph Klauer [a]

[a] Albert-Ludwigs-Universität Freiburg, Institut für Psychologie, D-79085 Freiburg, Germany
[b] Fachbereich Psychologie der Philipps-Universität Marburg, Gutenbergstraße 18, D-35037 Marburg, Germany
[c] University of California, Department of Psychology, One Shields Ave, Davis, CA 95616, USA

ARTICLE INFO

Keywords:
Social categorization
Social evaluations
Prejudice
Context effects

ABSTRACT

There is conflicting evidence for the extent to which spontaneous evaluations of outgroup members are influenced by social categorization. Contemporary models of evaluation like the APE model (Gawronski & Bodenhausen, 2006) imply that spontaneous evaluations are a function of the evaluative implications of the stereotypic associations of a target object and the context in which it is encountered. In this manuscript, we examine the question whether these models can also inform the debate on the link between social categorization and spontaneous evaluations: Specifically, we ask whether the link between categorization and evaluation is moderated by the subset of associations that is implied by a specific context. In two studies, we assessed the extent of categorization in conditions of the Who Said What? -paradigm with very different contextual implications for White versus Black people. This score was then correlated with the evaluative preference for White over Black people in an Evaluative Decision Task. In contexts in which it is an asset or advantage to be Black, there was less anti-Black bias the more an individual categorized by race. The reverse was true in contexts that are likely to activate negative racial stereotypes about Black people. A last confirmatory study with a larger sample did not replicate these correlations. The link between categorization and evaluation, and its contextual moderation, remains elusive.

If, hypothetically, at a boring department Christmas party,[1] the next party game would be to categorize social categorization researchers, the players might end up sorting those researchers into two different groups.[2] Those two groups would have common denominators. First, they would share the belief that social categorization, that is, the tendency to perceive the social world along distinct social categories, is a universal mental process (Allport, 1954): People routinely categorize others into commonly used categories, such as those pertaining to sex, age, and ethnicity, and this process appears to have a strong automatic or habitual component (e.g., Fiske & Neuberg, 1990). Second, the two groups would also agree that as a result of this categorization process, the target person can be perceived as part of the social group(s) to which the categorizer identifies as being a member (ingroup) or not (outgroup; e.g., Bodenhausen, Kang, & Peery, 2012). If members of an outgroup are routinely devalued, this is conventionally called prejudice, which is seen as a (possibly *the*) mental antecedent of intergroup conflict.

At this point, the groups' commonalities would end. One group of researchers emphasizes the role of categorization in prejudice. They conceptualize a causal (and graded) relationship between the two constructs, such that the (increased) perception of a target person as belonging to an outgroup is associated with (increased) prejudice against that person (see, e.g., Tajfel, 1982). The first evidence commonly cited in favor of this argument stems from minimal group paradigms (e.g., Tajfel, Billig, Bundy, & Flament, 1971), where participants are told that they are assigned into groups on some minimal, seemingly insignificant difference between them (for instance, their aesthetic

☆ This paper has been recommended for acceptance by Kristin Pauker.

* Corresponding author at: Fachbereich Psychologie der Philipps-Universität Marburg, Gutenbergstraße 18, D-35032 Marburg, Germany.
E-mail addresses: manuel.becker@psychologie.uni-freiburg.de, manuel.becker@uni-marburg.de (M. Becker), teige@uni-marburg.de (S. Teige-Mocigemba), jsherman@ucdavis.edu (J.W. Sherman), klauer@psychologie.uni-freiburg.de (K.C. Klauer).

[1] Please note that we can only speculate about the kinds of games people play at boring department Christmas parties, because none of our departments tend to host those.

[2] As per usual, many other categorization outcomes are conceivable, reflecting additional controversies in categorization research, e.g., Monroe et al. (2018).

https://doi.org/10.1016/j.jesp.2025.104842
Received 26 July 2023; Received in revised form 29 September 2025; Accepted 17 October 2025
Available online 6 November 2025
0022-1031/© 2025 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

preference for one painter over the other), while group assignment is actually random. Participants then typically favor other individuals from the group they have been assigned to over participants from the outgroup in self-reported liking or in their behavior (e.g., by favoring ingroup participants in the distribution of rewards). The first group of researchers also typically cites more indirect evidence. For instance, Todd et al. (2021) recently published a study in which attending to a category dimension other than race attenuated racial biases in deciding whether a target in a computer task holds a weapon or a tool. Likewise, Levine, Prosser, and Evans (2005) suggested that categorization influences intergroup helping behavior.

This group of researchers then typically infers from this assumption of a causal link between the two constructs that prejudice should be tackled by aiming at the participants' social categorization: If prejudice denotes the negative evaluation of a person based on his or her category membership, altering the categories along which others are perceived should have an impact on prejudice.

The other group of researchers is not so sure: In their view, social categorization may not contribute much to the emergence of prejudice above being a necessary condition for it, after all (for a review, cf., Park & Judd, 2005). These researchers can also refer to studies that back up their claim, some of which manipulate category salience directly. For instance, Deffenbacher, Park, Judd, and Correll (2009) have reported, among others, a study in which participants sorted picture portraits along a category dimension. This manipulation of category salience increased the extent of self-reported categorization, but there was no effect on the corresponding self-report prejudice measures.

Other studies in this line of research capitalize on the development and validation of an unobtrusive measurement procedure for social categorization, the so-called Who Said What? Paradigm (WSW paradigm; Taylor, Fiske, Etcoff, & Ruderman, 1978). In the WSW paradigm, participants follow a discussion in which speakers fall into different categories (e.g., Black and White speakers). After the discussion, the discussion statements are shown again along with the speakers, and the participants' task is to assign each statement to the speaker who made the statement. When assignment errors occur, they tend to be within-category errors, in which the statement is assigned to a wrong speaker from the actual speaker's category rather than between-category errors, in which the statement is assigned to a speaker of the wrong category. The error-difference measure contrasts the frequencies of within-category and between-category errors (with an appropriate chance correction) and is interpreted as a measure of the strength of social categorization with higher values indicating more within-category than between-category errors (Taylor et al., 1978).

Vescio, Judd, and Kwan (2004) took advantage of this development by comparing, in two seminal studies, simple and crossed conditions in the WSW paradigm. In the crossed conditions, race and gender were crossed so that target persons differed with respect to race (Black vs. White) and cutting across race, in gender (male vs. female) in Study 1, whereas relationship status (single or significant other) and hometown size (small or big) were crossed in Study 2. In the simple conditions, target persons differed in terms of only one such dimension (e.g., race), and their category membership with respect to the other dimension (e.g., gender) was held constant. In the WSW measure of social categorization, simple conditions elicited stronger categorization than crossed conditions, but self-report measures showed no evidence of stronger prejudice in simple conditions than in crossed conditions. Moreover, despite large sample sizes, there was little evidence of correlations between the WSW measure of social categorization and self-reported prejudice. The second group of researchers, that is skeptical of the role of categorization in prejudice, thus interprets these results as additional evidence in their favor.

These researchers thus suggest that factors other than the extent of categorization are decisive for whether and to what extent prejudice emerges. As a consequence, social categorization may not necessarily provide the best means to tackle prejudice: After all, we know that eyesight is a necessary condition for reading, but most interventions to foster reading do not specifically target improving people's eyesight.

If ideas on the role of social categorization in prejudice are so different, how could one attempt to increase the understanding between these two groups of social categorization researchers? One way to go about this could be to refocus on an old and common idea in social psychology, the power of the situation. Both social categorization and prejudice appear to be influenced not only by stable, person-related factors, but also by contextual factors. As described above, some amount of social categorization appears to be ubiquitous, especially along the most commonly used categories applied to human beings. Yet, it also appears to be influenced by a diverse set of cognitive, motivational and social factors, such as the accessibility of stereotype-related information, social status cues, in-group identification, relevance of the categorization dimension, task demands, or the presence of a common threat that is shared by the in- and outgroup (e.g., Castano, Yzerbyt, Bourguignon, & Seron, 2002; Correll, Guillermo, & Vogt, 2014; Eberhardt, Goff, Purdie, & Davies, 2004; Flade, Klar, & Imhoff, 2019; Freeman, Penner, Saperstein, Scheutz, & Ambady, 2011; Hugenberg, Young, Bernstein, & Sacco, 2010; Kurzban, Tooby, & Cosmides, 2001). There has also been much research interest in the malleability of prejudice. While some scholars see prejudice as a relatively stable entity, comparable to a trait-like psychological attribute (e.g., Fazio, 2007), others argue that such group evaluations are constructed on the spot, strongly influenced by contextual factors (e.g., Mitchell, Nosek, & Banaji, 2003; Schwarz, 2007). And, indeed, there is a sizeable literature on the contextual modulation of prejudice (e.g., Barden, Maddux, Petty, & Brewer, 2004; Wittenbrink, Judd, & Park, 2001; see Eagly & Diekman, 2005, and Gawronski & Sritharan, 2010, for reviews). In general, more egalitarian responses are found in contexts in which positive outgroup members and/or negative ingroup members are salient or in which outgroup members assume positive roles or are associated with less threat. However, there is also much evidence for the influence of stable person-related factors and person-situation-interactions in prejudice (Gawronski & Bodenhausen, 2017; Klauer & Becker, 2024; but see Payne, Vuletich, & Lundberg, 2017), and not for all context factors that were theoretically derived did researchers uncover convincing evidence (e.g., Joy-Gaba & Nosek, 2010; also see Alinor & Tinkler, 2021).

The interpretation of this complex relationship between person and situation in shaping prejudice becomes even more complicated when taking into account that prejudice can be conceptualized and has been operationalized as consisting of different components: The definition given above encompasses deliberate evaluations of groups and their members, a willingness to favor members of an ingroup over members of an outgroup, and spontaneous evaluations of outgroup members. Not all of these components of prejudice need to be equally malleable; and, circling back to the topic of interest, are likely similarly affected by categorization. Out of these components, social psychologists have been especially interested in spontaneous evaluations of outgroup members. This is because due to their speeded nature, the measurement outcome of procedures assessing these spontaneous evaluations might be typically less susceptible to participant control attempts that obscure true variation in prejudice (e.g., Gawronski & De Houwer, 2014; also see section "Measuring social evaluations on p. 8). Research on the link between categorization and prejudice by design renders accessible the intergroup context and its associated normative aspects, hence categorization-related conclusions based on measurement procedures assessing more deliberate components of prejudice might be even more likely to be different from conclusions based on measurement procedures less susceptible to participant control attempts. In this domain, scientists have thus to be especially careful not to conflate the different components of prejudice. We will restrict the scope of the present research and our inferences to spontaneous social evaluations, but will sometimes appeal to the larger literature on prejudice should specific insights on spontaneous social evaluations not exist yet.

Now that we have reviewed that both social categorization and social

evaluations are subject to contextual influences, we argue that the malleability of both constructs needs to be taken into account when examining the link between them. More specifically, we will propose that the strength and even sign of the correlation between categorization and social evaluations may be susceptible to contextual influences. To be able to do so, we will sketch the model for contextual influences on the relationship between categorization and social evaluations that guided our research, describe one experimental set-up in which the hypothesis that the link between categorization and social evaluations is context-dependent can be examined, and present evidence from two first studies conducted in this paradigm, followed by a confirmatory study with a large sample.

1. Contextual influences on the categorization-evaluation link

As referred to above, prejudice typically relates to a person's overall group evaluation. However, a person's knowledge structures about a social group contain more than a fundamental evaluation of the group. Instead, their knowledge structures contain a diverse set of attributes associated with a particular group (i.e., stereotypes), that are likely fundamentally intertwined with the group evaluation (Kurdi, Mann, Charlesworth, & Banaji, 2019; Phills, Hahn, & Gawronski, 2020). Importantly, the situational context appears to co-determine which subdomains of stereotype content are activated (Casper, Rothermund, & Wentura, 2010). For example, Schaller, Park, and Mueller (2003) argue that.

> There is an easy tendency to assume that stereotypes are monolithic cognitive structures characterized primarily by evaluative negativity (or positivity). This assumption leads easily to the further assumption that any one evaluatively loaded aspect of a stereotype (e.g., "Blacks are ignorant") can serve as a proxy for any other equally evaluative aspect of the stereotype ("Blacks are hostile"). Both assumptions are wrong. Stereotypes are highly dynamic knowledge structures composed of different elements that not only have evaluative implications but have different connotative implications as well. The results of the present studies suggest that we might best understand the influences on stereotypes and prejudice by attending carefully to the effects that specific variables have on specific domains of stereotype content (Schaller et al., 2003, p. 646).

As a consequence, one could assume that the link from categorization to social evaluations is at least partially mediated by the evaluative implications of the stereotypic associations evoked by the category. This reasoning is illustrated in Gawronski and Bodenhausen's (2006) Associative-Propositional Evaluation (APE) model. The APE model posits that mental concepts that represent an encountered stimulus are activated by a feature-matching mechanism. Activation from these concepts, in turn, spreads to other concepts linked to the target concept in memory. The authors give the example of features of a face activating the concept African-American, and this, in turn, activating a subset of associated mental attributes such as hostile. They assume that the extent to which the activated concepts have a positive or negative connotation in the aggregate is translated to a spontaneous affective response. Now, crucially, the APE model is built on the premise that the subset of mental attributes that is activated by the stimulus is a function of the overall configuration of input stimuli, that is, a combination of the target object itself and the context in which it is encountered. They elaborate:

> For example, encountering a Black person in a jazz bar may activate the stereotypical attribute musical, whereas the same Black person may activate the stereotypical attribute criminal if that person is encountered in a dark alley (Gawronski & Bodenhausen, 2006, p. 62).

Accordingly, the strength of the link between categorization and social evaluations may depend on which subset of attributes is salient. To the extent that a context activates favorable attributes associated with a social category, a larger extent of social categorization should lead to more favorable impressions of group members. In contrast, if the context activates unfavorable attributes, a larger extent of social categorization should lead to less favorable impressions of group members. In the following, we will report one research strategy that can put this model to a test. To be able to do so, we will first describe the literature on and constraints of measuring prejudice, as already foreshadowed above.

2. Measuring social evaluations

Spontaneous social evaluations are typically measured by so-called indirect measurement procedures. In the most common tasks, participants are often confronted with pictures of ingroup and outgroup members along with other stimuli (e.g., symbols; evaluatively polarized or stereotypically loaded words) and have to show some sort of, typically speeded, reaction to these stimuli. The underlying construct (i.e., the spontaneous evaluation of a social group) is then inferred from task performance.

For the present endeavor to examine context effects on the link between categorization and evaluation, not all of the most popular measures of social evaluations seem equally suitable. In the measurement procedure used most frequently for the assessment of social evaluations, the Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998), participants are required to sort target stimuli according to their social categories throughout the task, thereby bringing about high levels of category salience for all participants, possibly obscuring the effects of categorization and context. Considering the Affect Misattribution Procedure (Payne, Cheng, Govorun, & Stewart, 2005), there is a lively discussion on its automaticity features and thus on the question when and under which circumstances the AMP may allow the assessment of social evaluations in socially sensitive domains such as prejudice against minorities (e.g., Bar-Anan & Nosek, 2012; Hughes, Cummins, & Hussey, 2021; Mann, Cone, Heggeseth, & Ferguson, 2019; Payne et al., 2013; Teige-Mocigemba, Becker, Sherman, Reichardt, & Klauer, 2017).

For these reasons, we believe that in the present study context, the most suitable measure for spontaneous social evaluations is the Evaluative Decision Task (EDT; a measurement procedure from the family of Evaluative or Affective Priming Tasks; Fazio, Sanbonmatsu, Powell, & Kardes, 1986;Fazio, Jackson, Dunton, & Williams, 1995). In a standard version of the EDT measuring social group evaluations, briefly presented, task-irrelevant pictures of a member of either of two social categories are quickly succeeded by valenced target words, which are to be categorized in terms of their valence. Participants' performance in categorizing the target words is then analyzed as a function of the relationship between the preceding pictures (called primes) and the targets: So-called congruent trials, which include trials in which ingroup members are succeeded by positive words, and trials in which outgroup members are succeeded by negative words, are compared with the remainder of the trials, so-called incongruent trials, which include trials in which ingroup members are succeeded by negative words, and trials in which outgroup members are succeeded by positive words. Participants are instructed to respond quickly, but typically have ample time to do so, and the measurement outcome consists of the latency difference between congruent and incongruent trials with correct responses, termed *priming effects*. In line with the conventions from the social cognition literature (e.g., Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002), we will henceforth interpret the EDT measurement outcome as a reasonable proxy for social evaluations, defined as the measurement outcome of our indirect measurement procedure of choice for intergroup evaluations. Naturally, any of our claims are subject to reinterpretation with different, suitable measurement procedures and experimental designs.

3. The present research

The WSW paradigm is well-suited to investigate the influence of

evaluative implications of the context on the categorization-evaluation link not only because it is the most thoroughly validated measure for categorization (e.g., Klauer & Wegener, 1998), but also because its procedure naturally entails a rather strong context by virtue of participants having to follow a discussion between the measurement targets. By changing the conversation between participants (with random assignment of a large variety of statements to the speakers), the evaluative implications of the context in which a participant categorizes and evaluates can be readily manipulated. If one then wants to assess the role of categorization in spontaneous social evaluations, a straightforward strategy is thus to present the WSW measurement procedure for categorization followed by a measurement procedure for evaluations (in this case, with the EDT), and then correlate the two measurement outcomes.

The present research on a possible context manipulation borrows from the study by Vescio et al. (2004) reported above by implementing a similar experimental design. At the same time, it also differs from the Vescio et al. (2004) study in several regards: Firstly, we used an improved analysis strategy for the WSW measurement outcome, because systematic confounds and unsystematic noise in the traditional WSW error difference measure can otherwise affect results, as described in Klauer, Hölzenbein, Calanchini, and Sherman (2014). Secondly, we used the EDT as a measure for social evaluations and lastly, we did not compare simple vs. crossed categorization as a means for inducing differential levels of categorization.

Given that this line of research is situated in the tradition of a general link between categorization and prejudice, we first conducted a study with conditions between which categorization varies. We thus opted to find WSW statement pools that were known to differentially affect the overall level of social categorization (and, coincidentally, offer different evaluative implications). Fortunately, we were able to build on another previous line of research where the implications of a variety of statement pools have already been assessed. Klauer et al. (2014; Study 3) crossed race (Black versus White) and team membership (membership in one of two basketball teams) in a WSW paradigm. Different groups of participants watched discussions on different topics: on a conflict between the two teams concerning their last basketball match, on the situation at the university, or on racial issues in society. The different topics instigated different overall levels of social categorization in the original study. In particular, categorization by race increased from the team-conflict topic to the university topic to the race-relevant topic (Klauer et al., 2014, Study 3). The present Study 1 extended this design by adding an EDT as a measure of social evaluations of Blacks vs. Whites.

4. Study 1

Our main questions in Study 1 were whether individual levels of social categorization correlate with spontaneous social evaluations, holding topic group constant; and whether this relationship generalizes over the different contexts engendering different levels of categorization. We also tested as a subordinate hypothesis whether the different levels of social categorization in the different topic groups would be accompanied by corresponding differences in overall social evaluations between groups. As in the Klauer et al. (2014) study, we aimed at collecting data from 40 participants per group. Please note that for all studies reported in this manuscript, sample sizes were determined before any data analysis, and we report all measures, manipulations, and exclusions.

4.1. Method

Studies 1 and 2 always took place in individual laboratory cells, in which participants were seated at a distance of 60 cm from a 58 cm LCD TFT monitor (LG Flatron W2363D) with a refresh rate of 120 Hz. The studies were implemented on a Windows 7 system operating on a Fujitsu Esprimo P920/E85 personal computer. They were administered by a PsychoPy (Peirce, 2007) computer program. Participants received 3,50 € or partial course credit for taking part in the studies. The WSW part of Study 1 replicated Study 3 by Klauer et al. (2014). We therefore reproduce Klauer et al.'s (2014) description of the methods of this part in a shortened, but in some places almost verbatim version.

4.1.1. WSW procedure

There were three large pools of 192 statements each, one on conditions at the university (see Klauer, Ehrenberg, & Wegener, 2003; Klauer & Wegener, 1998), one on the merits of one's own basketball team and the weaknesses of the other team in respect to the last match of the two teams, and one on issues of racial equality in society.

Half of the university statements express satisfaction with different aspects of the conditions at the university, the other half express dissatisfaction. Critical and satisfied statements were constructed in similarly worded pairs so that, for each statement, there is a parallel one expressing the alternative attitude towards the same or a very similar aspect of the university. For example, the critical statement "The office hours are too short and there is hardly time for intensive talks" finds its more satisfied counterpart in "The office hours are long and there is sufficient time for intensive talks".

This same principle was followed for all statement pools. Thus, in the team statements, one member of each pair of parallel statements disparages the other team in claiming a certain weakness for that team relative to one's own team (e.g., "You were simply in worse shape physically than us"), whereas the other member claims a certain strength for one's own team relative to the other team (e.g., "We were simply in better shape physically than you") using almost the same words. In the race statements, one member of each parallel pair of statements claims the existence or persistence of race-based differences in how a person is treated or perceived in contemporary society (e.g., "Black and White persons doing the same job are often not paid the same") and the other member claims the absence of such differences using almost the same words (e.g., "Black and White persons doing the same job are usually paid the same").

Let us refer to statements of one polarity (i.e., for Topics 1 to 3, in order, critical, other-team disparaging, and claiming the persistence of race-based differences) as Type 1 statements, and to statements of the second kind as Type 2 statements. The statement pools (and the computer programs used in this study) can be found on the Open Science framework under the link https://osf.io/dk3zp/?view_only=7bc33ca589d045fa809ed811b38ddee1.

Each of twelve speakers made four statements that were randomly sampled (for each participant anew) with the restrictions that the speaker made two Type 1 and two Type 2 statements and that the two members of a statement pair never both occurred in the discussion. For the assignment phase, 48 additional distractor statements were sampled with these same restrictions (i.e., equal numbers of Type 1 and Type 2 statements, and the two statements of each pair never both sampled) from the statement pools from which the statements sampled for the discussion as well as their paired parallel statements had been removed. Note, in particular, that the two members of a statement pair were never both seen in the course of any participant's experimental session.

We used 12 pictures of young White and Black men taken from the official Internet site of the Mississippi Bulldogs track and field 2011 roster to represent the speakers. These showed the head and upper part of the body. The men in the photos wear red team jackets. Photo-editing software was used to remove the team label and the producer's label from the jackets and to create two parallel versions of each photo with jacket color changed to blue in one version and to a light green in the other version.

Each participant saw the six White and six Black speakers from these photos. Within race, three speakers were randomly sampled from the photos with green jackets, and three different speakers from the photos with blue jackets.

A participant's WSW session lasted about 15 min. Participants read that they were to follow a discussion of two rival basketball teams of the

same university and that their task was to form an impression of the group and the group members. They were told that members of one team wear green jackets, whereas members of the other team wear blue jackets, and that each team comprises both White and Black players. Participants were told that the two teams had competed against each other in the last season and were now engaged in a discussion.

Participants then watched the simulated discussion. Each contribution consisted of a written statement along with its speaker's photo. The photo was first presented for 1.5 s, followed by the presentation of the written statement along with the photo for 7.5 s and an inter-stimulus interval of 0.1 s. Speakers made statements in turns so that in a first round, all twelve discussants made their first statement followed by a second round of statements by each speaker, and so forth. The order of speakers was randomized in each round anew.

After observing the discussion of either the team-conflict topic, the university topic, or the race-relevant topic, participants were shown, in random order, all 48 discussion statements, respectively, mixed with 48 distractor statements. Their task was to decide first whether the statement had been previously presented by clicking on screen fields labeled "old" and "new". If they decided "old", the 12 speakers were presented in random order in two rows of six speakers, and participants were asked to assign the statement to its speaker by clicking on one of the 12 speakers. After a "new" decision or a person decision, the next statement appeared and was to be classified.

4.1.2. EDT

After the WSW task, participants completed the EDT. Target words were sampled from a set of 65 strongly polarized positive adjectives and a set of 65 strongly polarized negative adjectives that have been used in previous priming studies (e.g., Klauer, Teige-Mocigemba, & Spruyt, 2009). Prime stimuli were the portraits used in the WSW task.

A trial of the priming task began with the presentation of a fixation cross for 250 ms followed by the prime. After 200 ms, the prime was replaced by a target word. The target was taken off the screen upon the participant's response or 800 ms after target onset, whichever occurred first. The next trial started after the participant had responded. The intertrial interval was 500 ms. Participants were asked to respond correctly as quickly as possible while the target was on screen. They used the interior keys of two computer mice positioned to the left and right of them to respond. The mice were connected to the computer's parallel port, to improve the accuracy of response measurement relative to the use of standard mice or the keyboard (Voss, Leonhart, & Stahl, 2007). The response "negative" ("positive") was mapped on the left (right) mouse. In the case of a wrong response, the word "FEHLER!" [ERROR!] appeared, printed in red for 200 ms.

There were four blocks of priming trials. Each block began with four warm-up trials presenting two (randomly sampled) primes from each social category, one paired with a negative target, the other one with a positive target. The first block was a practice block presenting 24 additional trials and the remaining three blocks were experimental blocks, each presenting 48 additional trials. The additional trials comprised a balanced design with factors prime (12 pictures) and target valence (negative versus positive) within each block. Target words were sampled without replacement for each participant and block anew from the above-mentioned item pools. End-of-block feedback informed participants about the percentage of correct responses and mean latency in the preceding block.

A social evaluation score was computed following a protocol determined a priori that was in force for all studies in this paper. Specifically, trials with false responses were first discarded (thereby leaving out 9 % of the trials). Each participant's correct response latencies were then screened for outliers using Tukey's outlier criterion (Clark-Carter, 2004, Chap. 9); that is, for latencies 1.5 times the interquartile range below the first quartile or above the third quartile in the individual's distribution of correct response latencies. Outliers were discarded, excluding an additional 4 % of all trials. The remaining latencies were then log-transformed for the statistical analyses, though the reported descriptive statistics use the untransformed millisecond metric for ease of interpretation. For each participant, the measure of implicit prejudice was obtained by subtracting the means for (prejudice-congruent) trials with Black primes and negative targets as well as with White primes and positive targets from the means for (prejudice-incongruent) trials with Black primes and positive targets as well as with White primes and negative targets. Positive values thereby indicate more negative evaluations of Black relative to White people.

4.1.3. Participants

Participants were mostly students of a large German university with different majors who participated for partial course credit or a small monetary reward. Inclusion criteria were that participants had to be native German speakers without visual impairment. Furthermore, they were not to have participated in related experiments in our laboratory (using for example an EDT measuring prejudice), they were to be between 18 and 45 years old and, if they were psychology students, they were to be in their first semester (and so, had not completed courses on social-psychological theories of implicit prejudice and implicit measures). Data sets of seven participants were lost due to computer problems and two participants were excluded because they did not meet the inclusion criteria.

Participants were also excluded if they were extreme outliers according to Tukey's outlier criterion in terms of percent errors in the EDT and/or if percent errors exceeded 25 %. Two participants were excluded on this basis with mean percent errors exceeding 32 % (for the entire sample, $M = 9$ %, $SD = 7$). Furthermore, participants were excluded who were extreme outliers in mean response latency; four participants were excluded on this basis with mean latencies exceeding 890 ms (for the entire sample, $M = 586$ ms; $SD = 118$). This left 120 participants for the analyses of which 39, 40, and 41 were in the team-conflict, university, and race topic group, respectively. The hierarchical Bayesian analyses we conduct and report do not easily lend themselves to a conventional frequentist sensitivity analysis. Conceptually, power analyses answer frequentist questions that are not defined in a Bayesian framework, and there is no agreement on how to conceptualize something analogous to power in a Bayesian framework nor on how to estimate proposed analogues of power in a Bayesian hierarchical analysis (see De Santis & Gubbiotti, 2021; Kruschke, 2011, chap. 13; Schönbrodt & Wagenmakers, 2018, for examples of different proposals on how to define and assess sample-size requirements in Bayesian analyses). To meet journal conventions, we conducted a sensitivity power analysis for the overall correlation between categorization and evaluation holding topic group constant (indicating that the minimum effect size that could have been detected in a frequentist analysis, assuming 80 % power and an alpha level of 0.05, was $r = 0.25$), and for the separate correlations for each condition (indicating that the minimum effect size that could have been detected in a frequentist analysis was $r = 0.43$). To the extent that one wants to factor these results into the interpretation of the empirical results, one should take into account that the partial pooling involved in the hierarchical Bayesian analysis improves statistical efficiency so that conventional power analyses paint a conservative picture.

Mean age was 23.0 years ($SD = 5.2$); 65 % were female, 35 % male.[3] The raw data for all studies, including the data excluded from the analyses, and including some further demographic data, can be found alongside the study materials on the Open Science Framework under the link https://osf.io/dk3zp/?view_only=7bc33ca589d045fa809ed811b38ddee1.

[3] Note that we did not collect race or ethnicity information for either Study 1 or 2, but only a small minority of people in our participant pools typically identify as BIPoC, also cf. the demographic information for Study 3.

4.1.4. Data analysis

The WSW data were analyzed as in Klauer et al. (2014) using the mathematical model for WSW data with crossed categories proposed by Klauer et al. (2003). The model allows one to disentangle social categorization in the form of category memory from several other memory and guessing processes that also contribute to the error-difference measure and that may systematically bias it (Klauer & Wegener, 1998). The modeled processes and parameters are the following:

- Item memory I: the probability that a discussion statement is detected as an old statement or a distractor statement as a new statement.
- Person memory c: the probability that the statement's speaker is remembered for a discussion statement correctly detected as old.
- Category memory d(race), d(team | race), and d(team | not race): in order, the probabilities that race is remembered, team is remembered given that race is remembered, and team is remembered given that race is not remembered, with regard to the speaker of a discussion statement correctly detected as old, for which there is no person memory. This is one of several equivalent parameterizations of the two by two table for category memory, crossing the factors "race remembered" (yes vs. no) and "team remembered" (yes vs. no). Category memory for team can be computed as d(team) = d(team | race)d(race) + d(team | not race)(1 − d(race)).
- Item status guessing b: the probability that a statement is guessed to be old if there is no item memory for the statement.
- Category guessing a(Team 1) and a(Black): the probability of guessing that the speaker was from Team 1 rather than Team 2, and the probability of guessing that the speaker was Black rather than White, respectively, in guessing without memory. The probabilities to guess Team 2 and White are given by 1 − a(Team 1) and 1 − a (Black), respectively.

We used the same simple model as in Klauer et al. (2014), in which all memory parameters (I, c, and d) are set equal a priori across team and race as well as statement type, and only parameters a and b were allowed to vary as a function of statement type. The analysis was based on a Bayesian approach proposed by Klauer (2010) and Matzke, Dolan, Batchelder, and Wagenmakers (2015) to fit a multilevel extension of the model that treats participants and items as random factors (Judd, Westfall, & Kenny, 2012) for each model parameter, with topic condition as a fixed factor (consult Klauer et al., 2014, for a more detailed description; and Rouder & Lu, 2005, and Lee & Wagenmakers, 2013, for introductions to Bayesian hierarchical modeling).

This method naturally provides three model checks to assess goodness of fit (Klauer, 2010): Statistic *T1* summarizes how well the model accounts for the pattern of 58 observed assignment frequencies aggregated across items and participants within each condition, corresponding to the goodness-of-fit statistic *G2* used in the traditional approach (Batchelder & Riefer, 1999); *T2a* summarizes how well the model accounts for the variances and correlations of these frequencies computed across participants as units within each condition; *T2b* summarizes how well the model accounts for the variances and correlations of these frequencies computed across items as units within each condition. As can be seen in Table 1, the model checks for the baseline model were satisfactory in Study 1, although the results for *T2a* suggest that there is some room for improvement (large *p*-values indicate good fit).

Furthermore, the Bayesian approach yields highest density intervals (HDI) for the parameter estimates that can be interpreted much as classical confidence intervals, and it permits one-tailed hypothesis tests for equality between any two parameters. These hypothesis tests use the samples from the posterior distribution of the parameters calculated in the Bayesian analysis and estimate Bayesian *p* values defined as the posterior probability of the H_0 that differences between the two parameters are smaller than or equal to zero (see e.g., Casella & Berger, 1987).

Table 1
Model checks for the WSW baseline model.

Statistic	Observed	Expected	*p*
Study 1			
T_1	150.6	136.8	0.228
T_{2a}	6988	5699	0.029
T_{2b}	2362	2347	0.424
Study 2			
T_1	75.49	91.76	0.867
T_{2a}	3408	3488	0.598
T_{2b}	1607	1671	0.750
Study 3			
T_1	179.8	92.18	<0.001
T_{2a}	3727	3376	0.065
T_{2b}	4893	4495	0.073

4.2. Results

As a replication of Klauer et al. (2014) and to assess the magnitude of differences in categorization strength the different contexts promote, we first present the results for the measure of categorization: In the original study, categorization by race as tapped by the category memory parameter d(race) increased from the team-conflict topic to the university topic to the race topic. Next, and as a test of our subordinate hypothesis, we consider the outcome of the measure of social evaluations to see whether the differences between groups in the strength of categorization by race are reflected in it. Finally, and central to the present investigation, we report the correlations between social categorization by race on the one hand and the measure of spontaneous social evaluations on the other hand. We report one-tailed *p*-values where our hypotheses were directed (i.e., the replication analyses), as detailed below.

The error bars give the 95 % HDI intervals. On the y axis, category memory parameters d(team) and d(race) as our measures of categorization are depicted.

4.2.1. Categorization

Figure 1 shows the categorization by team membership, d(team), and by race, d(race), as a function of discussion topic along with the 95 % HDIs. Parameter estimates and HDIs for the other model parameters are shown in the online supplemental materials. As in Klauer et al. (2014), race categorization increased from the team-conflict topic to the university topic to the race topic, whereas team categorization simultaneously decreased and was weaker overall. Reflecting the present focus on race categorization, we compared race categorization (1) in the team-conflict topic versus university topic conditions and (2) in the university topic versus race topic conditions. There was evidence for increases in race categorization in both comparisons, but only weakly so for the second contrast comparing university topic and race topic, with one-tailed Bayesian p − values of $p < .001$ and $p = .063$, respectively.

4.2.2. Social evaluations

As can be seen in Table 2, mean priming effects increased along with social categorization by race. There was a small pro-Black bias of −3 ms in the team-conflict group, almost no bias (a 1 ms priming effect) for the university topic, and a pro-White bias of 8 ms in the race-topic group. The differences between these priming effects were significant, $F(2, 117) = 3.76$, $p = .026$, $\eta_p^2 = 0.06$.[4] Only the 8 ms effect in the race-topic group was significantly different from zero, $t(40) = 2.69$, $p = .010$, $d_z =$

[4] For the sake of comparability between studies, note that this translates to a significant difference between the race-topic and the team-conflict condition, $t(78) = 2.755$, $p = .007$, $d = 0.62$ in a post-hoc test.

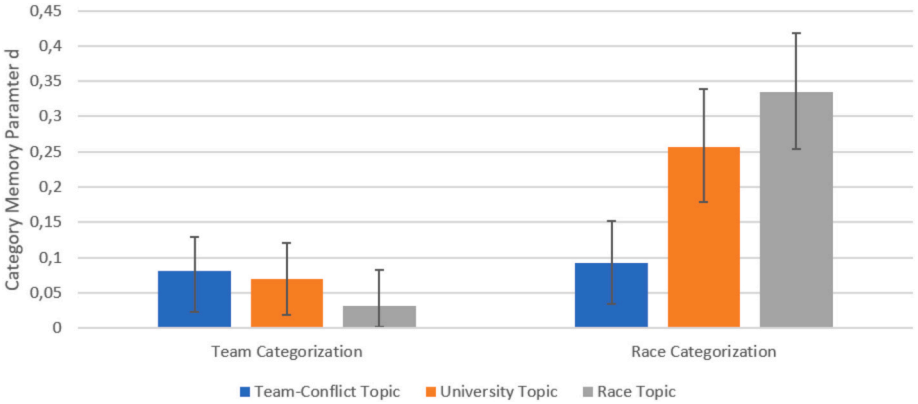

Fig. 1. Categorization by team membership and race in Study 1.

Table 2
Mean response latencies in ms as a function of prime type and target valence, and priming effects for each condition of studies 1 to 3.

	Black Prime				White Prime					
	Neg. Tar.		Pos. Tar.		Neg. Tar.		Pos. Tar.		Priming	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Study 1										
High team rel.	566	72	541	75	560	71	540	75	−2.8	16.5
Low rel.	567	73	539	66	569	69	539	68	0.8	19.3
High race rel.	568	59	545	57	577	60	537	51	8.2	20.5
Study 2										
Leisure activities	560	58	535	55	559	54	532	51	0.9	23.2
Race topic	547	66	518	49	550	64	616	48	2.9	18.2
Study 3										
Advantaged-Blacks	545	66	548	68	524	61	524	60	1.6	13.3
Disadvantaged-Blacks	556	62	538	62	558	62	535	61	2.6	14.3

Neg. = negative; Pos. = positive; Tar. = target; Rel. = relevance.

0.42. The effect in the team-conflict condition, indicating pro-Black bias, was in the unexpected direction and non-significant, $t(38) = -1.10$, $p = .279$, $d_z = -0.18$. The effect in the university topic condition also was non-significant, $t(39) = 0.50$, $p = .620$, $d_z = 0.08$.

4.2.3. Correlational analyses

The correlational analyses were based on an extension of the above baseline model: The z − standardized EDT score was entered as a predictor into the equation for an individual's model parameters (see online supplemental materials for additional information). The analysis using this model yields a regression weight β for each parameter (and a HDI around it), with topic group controlled for. There was no evidence for an overall relationship between social categorization by race and social evaluations (over and above the just-described group-level differences in priming scores and categorization), $\beta = -0.005$, 95 % HDI = (−0.16, 0.16), with a Bayesian p − value of 0.524 for the hypothesis $\beta > 0$ (H1). In subsequent analyses, we fitted the model to each topic group separately. In the team-conflict group, there was a tendency for a reversed regression coefficient, $\beta = -0.36$, 95 % HDI (−0.84, 0.09), indicating more positive evaluations of Blacks the stronger the categorization by race. There was no evidence for a relationship in the university-topic group, $\beta = -0.06$, 95 % HDI (−0.31, 0.18), and a tendency for a correlation in the expected direction with more negative evaluations of Blacks the stronger the categorization by race, $\beta = 0.14$, 95 % HDI (−0.19, 0.50), in the race topic group. The resulting correlations and their distributions are shown in Fig. 2 for each group.

The groupwise results suggest — as do the group means of the EDT — that the relationship between categorization and evaluations is reversed

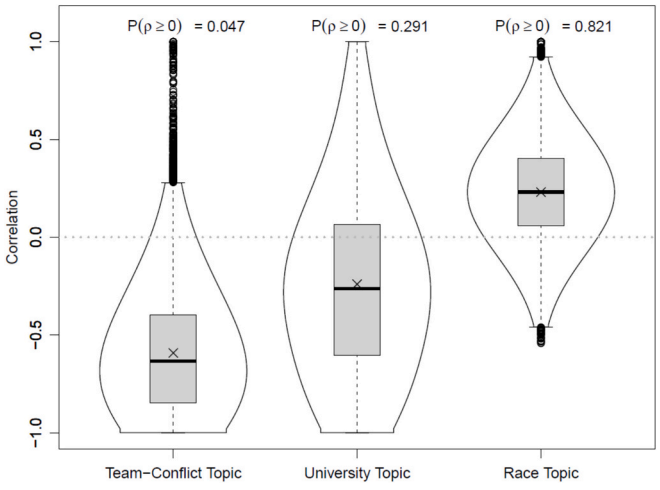

Fig. 2. Relationship between social categorization and spontaneous evaluations in Study 1.
Violin plots of the posterior distribution of the correlation between social categorization by race and evaluation score in Study 1. The boxes give the interquartile range; the crosses the means; the bold horizontal lines in the boxes the medians; overlaid to the left and right are kernel density estimates of the posterior distribution.

in the team-conflict topic group, small to non-existent in the university-topic group, and positive in the race-topic group. To examine this possibility further, we fitted a model that additionally included the interaction of EDT scores and group as predictor in order to test for moderation of the correlational link by topic group. The interaction was coded in terms of two contrasts, of which the first, C1, contrasts the team-conflict and the race-topic group. In this analysis, there was evidence for a credible difference between the regression weights of the team-conflict and the race-topic group, C1 = 0. 36, 95 % HDI (0.02, 0.73), with the 95 % HDI for the contrast excluding the zero point of no difference.[5]

4.3. Discussion

Study 1 replicated the pattern found in Klauer et al.'s (2014) Study 3 for categorization by race and team. Race categorization was highest when the topic of race was made salient. More importantly, there were overall group differences in the social evaluation measure as a function of discussion topic, with more categorization going along with more bias. The effect was, however, small ($\eta_p^2 = 0.06$). In line with the conclusions from Vescio et al. (2004), there was absolutely no indication for an overall correlation between categorization and social evaluations. As described above, from the perspective of conventional models of the link between categorization and prejudice, this is surprising.

Instead, the present findings suggest a more nuanced result in which the correlational relationship was moderated by discussion topic. More categorization by race tended to go along with less pro-White bias in the team-conflict topic and with more pro-White bias in the race-topic group. As mentioned above, our aim for the first study was to investigate the generalizability of the link between categorization and social evaluations over several contexts that differentially encourage categorization. A model in line with the conclusions by Vescio et al. (2004) does not account for the present pattern of results, nor does an extension of the conventional models that naively assumes that the relationship between categorization and prejudice is a function of the overall level of categorization. However, viewed from the perspective of our model sketched above that assumes that the strength and sign of the link between categorization and social evaluations depends on which subset of attributes is salient, these results are rather straight-forward to explain. Most notably, in the race-topic condition, the discussion highlighted intergroup conflict whereas in the team-conflict condition, the discussion made salient athletics in general and basketball in particular. There is ample evidence that Black people are strongly associated with athletics in general and basketball, in particular, and that this stereotype is clearly positively valenced (see, e.g., Eberhardt et al., 2004, for the US-American context). The reasoning behind the above-mentioned model translates into the assumption that the team-conflict topic, being related to basketball, would tend to cue positive aspects of the stereotype ("Black people are good athletes") with respect to which it is an asset to be Black. In contrast, the race topic would tend to cue intergroup conflict and negative aspects (e.g., "Black people are less intelligent") with respect to which it is disadvantageous to be categorized as Black. In consequence, categorizing a Black person by race will tend to activate stereotypic associations with positive implications for the team-conflict topic and stereotypic associations with negative implications for the race topic. Taken together, stronger categorization by race should go along with less pro-White bias in the team-conflict topic group and with more pro-White bias in the race-topic group. Thus, these results are well accounted for by the model introduced above that suggests that the evaluative implications of the situation are decisive for the link between categorization and evaluation.

Study 2 followed up on this idea using modified statement pools designed to bring out the intriguing pattern of correlations more strongly. Specifically, the team-conflict topic was replaced by a leisure-activities topic discussing such activities (e.g., related to dance, sports, or music) in which Black people are stereotypically assumed to do better than Whites and thus, for which it is an asset to be Black. Note, however, that none of the statements made any reference to Black or White persons. In turn, the race topic was edited to remove contents referring to leisure activities, replacing about 25 % of the statements. Our hypothesis was that this would allow us to replicate and accentuate the correlational pattern found in Study 1.

5. Study 2

Previous research has identified dimensions with reliable positive stereotypic beliefs about African-Americans in the US (e.g., Czopp & Monteith, 2006; Madon et al., 2001) that are known in Germany (Ege & Hurley, 2015) and formed the basis for the new statement pool of the leisure activities topic in Study 2. Our predictions (which we now tested with one-tailed tests) were that we would find (a) a negative correlation between the strength of categorization by race (WSW measure) and social evaluations (EDT) in the group with the leisure-activities topic and (b) a positive correlation in the group with the race topic. Based on our findings and considerations from Study 1, we expected - as a subordinate hypothesis - that there would be a substantial preference for White over Black people in the condition with the race topic. We had no a priori prediction (and thus perform a two-tailed test) for the EDT outcome in the leisure topic group, as it is difficult to predict how much the favorable context balances out any overall negative group evaluations the participants might hold against Blacks in principle. We also had no a priori prediction for the effect of discussion topic on mean level of social categorization, as the statements from the leisure condition very obviously include topics from a variety of issues that are associated with Black people.

5.1. Method

The procedures of the WSW paradigm and the EDT followed those of Study 1, as did the groupwise target sample sizes. The university topic was no longer used and the team-conflict topic was replaced by a leisure-activities topic. The race topic statements were edited and, in part, replaced by new statements that did not refer to ability differences in leisure activities. As before, topic was a between-participants factor, but it now had only two levels, leisure activities and racial issues.

5.1.1. WSW statement pools

The statement pool for leisure activities was constructed as were the statement pools in Study 1. Thus, there were 96 statement pairs addressing how different people fare in various leisure contexts, like dancing, expressing themselves through music, doing sports, or socializing. Again, the two members of a statement pair addressed the same issue using almost the same words. The first member of each pair (henceforth, Type 1 statements) suggested that there are individual differences in the ability to perform well in an activity related to the subtopic, in some cases offering or implying essentialist explanations thereof (e.g., "It is said that only few people can learn to play basketball really skillfully"), whereas the second member of each pair denied the existence of individual differences or suggested that, where such differences exist, they can be overcome by training and practice using almost the same words (e.g., "It is said that most people can learn to play basketball really skillfully"). The statement pools are also accessible at the Online Repository.

[5] Note that for the second contrast that compares the university-topic group with the other two groups, which was computed for methodological reasons only and is thus not related to any substantial theory-related question, there was no evidence for a credible difference between the regression weights, C2 = 0.03, 95 % HDI (−0.28, 0.30).

5.1.2. EDT

Response latencies were preprocessed as in Study 1. Thus, trials with wrong responses (8 %) were excluded as were trials with outlying response latencies (4 %).

5.1.3. Participants

Participants were mostly students of a large German university with different majors who participated for partial course credit or a small monetary reward. Four participants were excluded because they did not meet inclusion criteria (see Study 1 methods). One additional participant was excluded prior to any data analysis because the participant indicated that they did not read the instructions.

In preparatory data analyses, further participants were excluded as in the previous studies if they were extreme outliers according to Tukey's outlier criterion in terms of percent errors in the EDT and/or if percent errors exceeded 25 %. Four participants were excluded on this basis with mean percent errors exceeding 26 % (for the entire sample M = 10 %, SD = 8). Furthermore, one participant was excluded as an extreme outlier in mean response latency in the EDT with a mean latency of 980 ms (for the entire sample M = 550 ms; SD = 91). This left 80 participants for the analyses of which 39 were in the leisure-activities group and 41 in the race topic group. Mean age for the participants was 23.8 years (SD = 5.1), 66 % were female, 34 % male. We again conducted a sensitivity power analysis for the separate correlations for each condition (now for a one-tailed test, indicating that the minimum effect size that could have been detected in a frequentist analysis was r = 0.39).

5.2. Results

As in Study 1, we present separate results for the measures of categorization and prejudice before the correlational results. Again, model fit was satisfactory (see Table 1).

5.2.1. Categorization

The same baseline model was fitted as in Study 1. Fig. 3 shows the categorization by team membership, d(team), and by race, d(race), as a function of discussion topic along with the 95 % HDIs. Parameter estimates and HDIs for the other model parameters are shown in the online supplemental materials. As can be seen, there was little categorization by team membership and substantial categorization by race, with little evidence for differences as a function of discussion topic. In particular, the 95 % HDI of the difference between the two groups in race categorization, $d_{\text{race topic}}(\text{race}) - d_{\text{leisure activities}}(\text{race})$, was (−0.15, 0.07) and contains zero as a credible value.

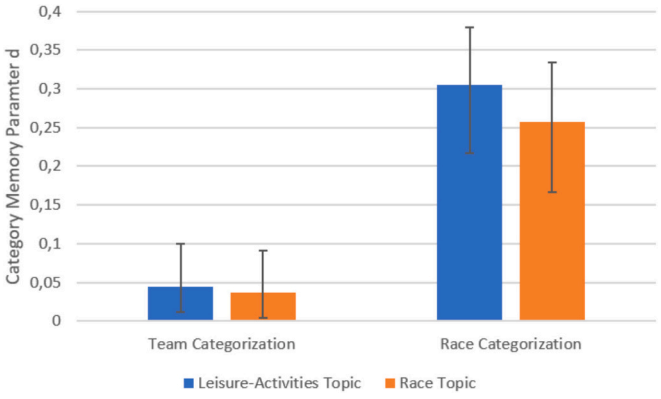

Fig. 3. Categorization by team membership and race in Study 2.
The error bars give the 95 % HDI intervals. On the y axis, category memory parameters d(team) and d(race) as our measures of categorization are depicted.

5.2.2. Social evaluations

As can be seen in Table 2, the mean priming effects were 1 ms and 3 ms in the groups with leisure-activities topic and race topic, respectively. In contrast to Study 1, neither of them was significant when tested individually (for the race topic group, $t(40) = 1.00$, one-tailed $p = .162$, $d_z = 0.16$; for the leisure-activities group, $t(38) = 0.34$, two-tailed $p = .739$, $d_z = 0.05$). Nor was their difference significant, $t(78) = 0.34$, two-tailed $p = .738$, $d = 0.08$.

5.2.3. Correlational analyses

We fitted the model with the z − standardized social evaluation measure as predictor to each group. In the leisure-activities group, the regression coefficient was $\beta = -0.29$, 95 % HDI (−0.66,−0.026), indicating that more categorization went along with less bias. Indeed, the one-tailed Bayesian p − value for $\beta < 0$ (H1) was $p = .012$. In contrast, in the race-topic group, the regression coefficient was $\beta = 0.25$, 95 % HDI (−0.01, 0.510), indicating that more categorization went along with more bias. And, indeed, the one-tailed Bayesian p − value for $\beta > 0$ (H1) was $p = .024$. Fig. 4 shows the resulting correlations and their distributions.

Not surprisingly, in a joint analysis of both groups with the social evaluation score, group, and the interaction of group and prejudice as predictors, there was substantial evidence for the regression coefficient being larger in the race-topic group than in the leisure-activities-topic group, $\Delta\beta = 0.52$, 95 % HDI (0.20, 0.85): The one-tailed Bayesian p − value for $\Delta\beta > 0$ (H1) was $p = .004$.

5.3. Discussion

Study 2 more pointedly tested the hypothesis that the relationship between categorization and spontaneous evaluation is a function of the evaluative implications of the overall configuration of inputs in a given situation. In a context in which it is an asset or advantage to be Black, there was less anti-Black bias the more an individual categorized by race. Conversely, in contexts that are likely to activate negative racial stereotypes, there was more anti-Black bias the more an individual categorized by race. In other words, context moderated the correlational link between categorization and evaluation. This moderation was so pronounced that it changed the sign of the correlation from significantly

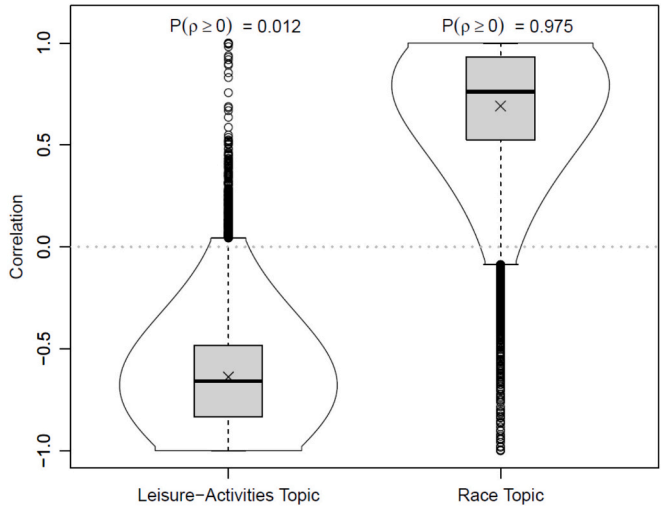

Fig. 4. Relationship between social categorization and spontaneous evaluations in Study 2.
Violin plots of the posterior distribution of the correlation between social categorization by race and evaluation score in Study 2. The boxes give the interquartile range; the crosses the means; the bold horizontal lines in the boxes the medians; overlaid to the left and right are kernel density estimates of the posterior distribution.

positive to significantly negative.

Unlike in Study 1, there was no significant evidence for anti-Black bias in the race-topic group. One difference between the race-topic groups in the two studies was that a number of statements were replaced (as detailed above) in moving from Study 1 to Study 2. However, it is not clear to us how this could cause a difference in social evaluations between the two studies. In fact, a joint analysis of the data from the race-topic groups of Studies 1 and 2 revealed a significant evaluative-priming effect, $t(81) = 2.63$, one-tailed $p = .005$, $d_z = 0.29$, with little evidence for an effect of study, $t(80) = 1.271$, $p = .208$, so that, overall, there is evidence for anti-Black bias in the groups with race topic. Still, the present study introduces some ambiguity. A contextual moderation without effects on the mean level of prejudice is not unheard of. For example, Maddux et al. (2005, Studies 1 & 3) report context-dependent moderation of a correlational link between motivation to control prejudiced reactions and implicit prejudice without effects of context on the mean level of implicit prejudice. At the same time and given the sizeable literature on the contextual moderation of prejudice, an absence of a main effect would need explaining if the contexts are as unambiguously evaluatively polarized as we suggest. In this light, the most likely explanation for the mixed results is variability due to the admittedly rather moderate sample sizes. Naturally, this needs to be further empirically corroborated. This is one of the reasons why we propose to conduct a registered Study 3 to further examine the effects.

In creating the leisure condition, we focused on designing a context in which Blacks are typically seen as advantaged, as described above. This is much easier if Black and White people are not explicitly mentioned in the statements, because stating that Black people are (naturally) better in the domains that are positively associated with them is commonly seen as a form of racism, and such statements might thus not create an unambiguously advantageous context for Black people (e.g., Kay et al., 2013). However, as a consequence of the fact that the "race" condition does mention race in every statement, the two conditions differ not only in terms of their evaluative implications, but also in whether race is explicitly mentioned in them. The second reason for conducting a registered Study 3 is thus to find out whether the present results hold even if race is not explicitly mentioned in either condition. Given that the parameter for category memory was very similar for the two conditions, a context with a salient advantage for either race appears to be sufficient to bring about high levels of categorization. Thus, it seems likely that abstaining from mentioning race in the condition that is disadvantageous for Blacks comes at little cost. To mark this change, we will call this condition the "disadvantaged-Blacks" condition from now on, whereas the other will be equivalently called the "advantaged-Blacks" condition. A plausible objection by a reviewer also prompted us to change five pairs of family-related statements in the advantaged-Blacks condition which, in the version of Study 2, were phrased in a way that might for some people trigger negative associations for Black people.

6. Preregistered study 3

The magnitude of the contextual moderation of the link between categorization and social evaluations demonstrated in the above studies is intriguing. From our perspective, the present findings could both add another layer to the debate on the link between categorization and prejudice and inform our understanding of the role of the situation in this relationship. The goal of the present work was to provide a firm and reliable answer to the question of whether different contexts (as operationalized by WSW conversation topics) can influence the direction of the link between categorization and spontaneous evaluation (as measured by the WSW and EDT, respectively). We intended to take this as a starting point for research on a more precise model for the link between categorization and social evaluation and its contextual modulation.

However, in the past, much time and resources have been spent chasing empirical phenomena that turned out to be much less robust than originally assumed, which lead us to our third reason for conducting Study 3: A rather direct and confirmatory replication of Study 2 with a larger sample could provide a better basis for trying to uncover and render more precisely the mechanisms behind the contextual modulation of the link between categorization and spontaneous social evaluations. As mentioned above, this replication not only allows us to investigate the robustness of the correlational differences, but also enables us to again test for the presence of anti-Black bias in contexts with negative implications for Black people, and rule out the possibility that effects are largely driven by the fact that Blacks and Whites are explicitly mentioned in only one condition.

6.1. Method

6.1.1. Participants

In light of the Bayesian nature of our analyses, it is difficult to perform a power analysis for the difference between correlations, see above. However, as described above, a second goal of the present study was to reassess whether the race context reliably leads to a devaluation of Black people in the EDT. Our joint analysis of the first two studies produced an average effect size of $dz = 0.29$ for the race condition. To achieve a power of 0.9 for an effect of this magnitude, $N = 104$ participants need to be tested, which is, conveniently, more than 2.5 times the size of the original participant samples (which is a benchmark for frequentist replications, Simonsohn, 2015). We therefore proposed to collect data until a sample size of $n = 104$ (after exclusions) has been reached for each condition, for a total of at least $N = 208$ participants. We opted to exclude participants by the same standards as described for the first two studies. In addition, we asked the participants a question about their self-identification as White or BIPoC, to make sure that Black people are indeed members of an outgroup for the participants. Our preregistered analyses rely on the full sample of participants that identify as White, but we describe the results for the full sample, as well. In addition to the laboratory where the first two studies were conducted, the third study was conducted at a second large German university.

The data of $n = 3$ participants were lost due to technical problems, $n = 1$ participant erroneously took part twice in the experiment, and the data of $n = 40$ participants were not used as they did not meet all of the inclusion criteria (native German speakers between 18 and 45 years old without visual impairment who are not advanced psychology students). Another $n = 3$ participants were excluded due to their performance in the EDT. Furthermore, we had preregistered an analysis of $n = 104$ participants per condition, but due to logistical reasons, collected some more data than preregistered. Consequently, the last $n = 11$ participants were excluded as excess participants in the disadvantaged-Blacks condition, and the last $n = 18$ participants were excluded as excess participants in the advantaged-Blacks condition. In a footnote, we report the results of an analysis including all participants fulfilling the inclusion criteria. Note that for this exploratory analysis, we excluded $n = 2$ further participants that were outliers in terms of item memory in the WSW procedure, an exclusion criterion we had not previously applied nor preregistered.

Additionally, and as preregistered, the data of $n = 7$ BIPoC

participants were entered into a joint analysis with the $N = 208$ participants from the main analysis. Given that so few participants identified as BIPoC, we did not analyze their data separately, as we had originally preregistered. In the sample for the main analysis, 75 % of participants identified as female (and 25 % as male), and 9 % were Psychology students. The participants' mean age was $M = 22.7$ years ($SD = 4.0$). $N = 54$ (26 %) of these participants took part in the laboratory used for the first two studies, while the remainder participated at the second university.

6.1.2. Procedure and materials

For the first university, the procedural details of the third study exactly mirrored our descriptions of Study 2, but the technical equipment was different: The single laboratory cells were equipped with HP EliteDesk 800G5TWR personal computers on which a Windows 11 system is operating and 60 cm LCD TFT monitors (LG 24GL600F) with a refresh rate of 144 Hz. For the second university, the single laboratory cells were equipped with Hyundai ITMC Pentino H series personal computers on which a Windows 10 system is operating and 61 cm LCD monitors (BenQ XL2430-B) with a refresh rate of 144 Hz. At both places, participants received 8,00 € or partial course credit.

As for the measurement procedures, we closely followed Study 2, with the exception of the altered statements in the WSW task reported above. At the end of the four blocks of EDT trials that were shown in the first two studies, four further blocks of trials followed. We report results from the first half of the task as a rather direct replication in a Footnote, and results for the full length of the task in the main body of the text. Following classical test theory, this increase in trials could boost the task's reliability, which might in turn further reinforce correlations. Apart from this and the explicit exclusion of Non-White participants mentioned above, we adhered to the analysis protocol from Study 2 for all analyses.

6.2. Results

A summary of our analysis preregistration can also be found at https://aspredicted.org/tmdm-sz26.pdf. As for the previous studies, we report model fit estimates for the main analysis of Study 3 in Table 1. The T statistics for the size of deviations of the data from the model predictions were in the same order of magnitude as in the previous studies, but as might be expected given the much-increased test power the associated Bayesian p values were less satisfactory than for the previous studies.

6.2.1. Categorization

The same baseline model was fitted as in the previous studies. In Fig. 5, categorization by team membership, d(team), and by race, d (race), is shown as a function of discussion topic along with the 95 % HDIs. While there was again little categorization by team membership and substantial categorization by race, in this study, categorization by race differed substantially as a function of discussion topic. The median parameter value for d(race) was estimated as higher for the disadvantaged-Blacks condition (0.31) than for the advantaged-Blacks condition (0.17). Consequently, the parameter difference between the two groups in race categorization, $d_{\text{disadvantaged-Blacks}}(\text{race}) - d_{\text{advantaged-Blacks}}(\text{race})$ was estimated as 0.13, and the 95 % HDI of this difference (0.05, 0.21) did not contain zero as a credible value.

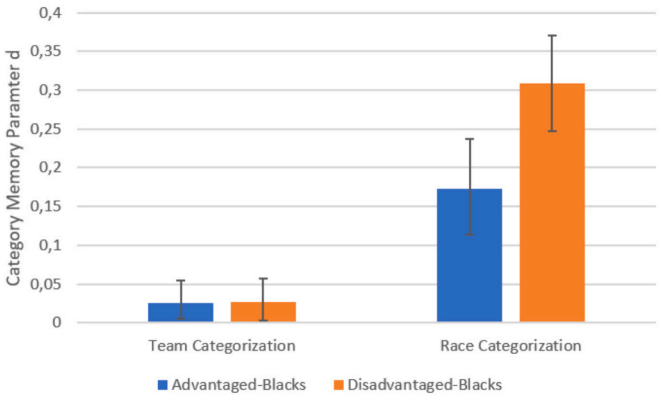

Fig. 5. Categorization by team membership and race in Study 3.
The error bars give the 95 % HDI intervals. On the y axis, category memory parameters d(team) and d(race) as our measures of categorization are depicted.

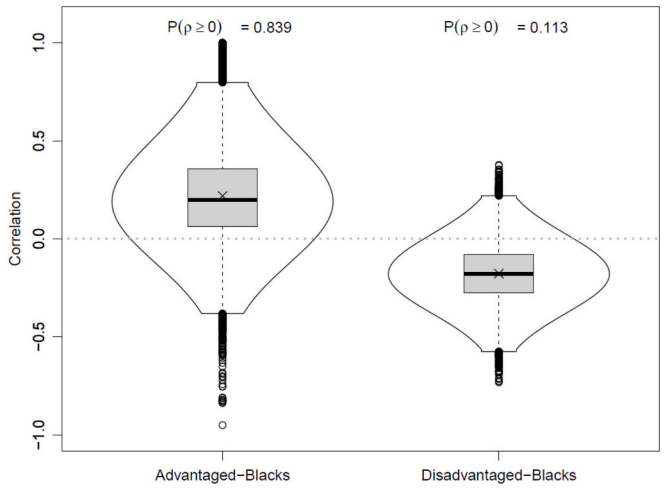

Fig. 6. Relationship between social categorization and spontaneous evaluations in Study 3.
Violin plots of the posterior distribution of the correlation between social categorization by race and evaluation score in Study 3. The boxes give the interquartile range; the crosses the means; the bold horizontal lines in the boxes the medians; overlaid to the left and right are kernel density estimates of the posterior distribution.

6.2.2. Social evaluations

The mean priming effects were $M = 2$ ms (advantaged-Blacks condition) and $M = 3$ ms (disadvantaged-Blacks condition, s. Table 2). In the advantaged-Blacks condition, the priming effect did not reach significance when tested individually, $t(103) = 1.064$, two-tailed $p = .290$, $d_z = 0.10$, contrary to the disadvantaged-Blacks condition, $t(103) = 1.859$, one-tailed $p = .033$, $d_z = 0.18$. The difference between these two effects did not reach significance, $t(206) = 0.604$, two-tailed $p = .547$, $d = 0.08$.

6.2.3. Correlational analyses

We again fitted the model with the z − standardized social evaluation measure as predictor to each group. In the advantaged-Blacks condition, the regression coefficient was $\beta = 0.10$, 95 % HDI (−0.11,0.29), with the 95 % HDI containing 0 as a credible value. Descriptively and contrary to our hypotheses, more categorization went along with more bias in this condition. In the disadvantaged-Blacks condition, the regression coefficient was $\beta = -0.09$, 95 % HDI (−0.24, 0.06), with the 95 % HDI again containing 0 as a credible value. Descriptively and contrary to our hypotheses, more categorization went along with less bias in this condition. Fig. 6 shows the resulting correlations and their distributions.

We then again conducted a joint analysis of both conditions, with condition assignment, the social evaluation score, and the interaction between them as predictors. As foreshadowed above, the regression cofficient was descriptively even larger in the advantaged-Blacks

condition, $\Delta\beta = -0.20$, but the 95 % HDI (−0.04, 0.44) contained zero as a credible value.[6]

6.3. Discussion

Study 3 was a close conceptual replication of Study 2, with the exception a) that the condition in which it is a detriment to be Black was altered such that this condition also does not in any way mention race anymore, and the exception b) that a few family-related statements from the condition in which it is an asset to be Black were altered such as to avoid possible negative associations they might entail. The study was conducted with a much larger sample than the two previous studies. Surprisingly, the crucial pattern of correlations from these studies was not reproduced at all; if anything, the regression coefficients in this study and in Study 2 have opposite signs. There was no evidence that, for instance, in a context in which it is a detriment to be Black, more categorization is associated with more negative evaluations of Blacks.

Another noticeable deviation of the present results from those of Study 2 is the diminished estimate for racial categorization in the advantaged-Blacks condition, while the substantially revised disadvantaged-Blacks condition showed categorization estimates similar to those observed in all prior conditions that render racial categories salient. More in line with the previous studies' results is the significant, but small anti-Black bias in the EDT. The significant effect is reassuring for an interpretation of this condition as a negative context for Blacks, even though the size of the effect leaves room for skepticism.

7. General discussion

The psychological literature is filled with studies that make contradictory statements on the role of categorization in prejudice: Some suggest that more categorization engenders more evaluations along category lines, while others suggest that the role of categorization is negligible. The first two studies presented here showed a consistent pattern of results that generated an intriguing hypothesis: That the link between categorization and evaluations might be moderated by the evaluative implications of the context in which a person is perceived. In light of the results of a well-powered third study, this hypothesis now has little empirical support. The present studies thus join the large number of studies with ambiguous results on the role of categorization in prejudice.

Why do the results differ between studies? The most parsimonious explanation is that the first two studies with their smaller sample sizes just randomly produced analogous correlational patterns (that, coincidentally, meshed well with established theoretical frameworks). Another possibility is that there are systematic differences between the samples of participants. Participants from Study 3 came from a different cohort of students and in part from a different university. Finally, both context conditions were altered for the third study, and the conceptually different results might be caused by these changes. It is difficult without further data to discriminate these possibilities, but we note, regarding the third possibility, that very few changes were made for the advantaged-Blacks condition for Study 3 (less than 5 % of the statements of Study 2 were altered, meaning that participants would on average see only 2 to 3 different statements out of 48 statements in the discussion phase). Yet, results for this condition also differed substantially from those of Study 2 in terms of, in particular, social categorization (compare Figs. 3 and 5) and in terms of the regression coefficients. This suggests that the changes in the context condition alone cannot account for the different results.

However, the disadvantaged-Blacks condition from Study 3 and the corresponding race-topic condition from Study 2 differed more substantially in terms of the used statements. Although the mean level of social categorization was comparable between the two conditions, item memory (parameter *I*) was considerably stronger in the disadvantaged-Blacks condition from Study 3 than in the race-topic condition from Study 2 so much so that it surpassed item memory in the advantaged-Blacks condition from Study 3: $\Delta I = 0.14$, *95 % HDI* (0.08, 0.2), $p < .001$ (s. the Online Appendix). Again, it is difficult to say whether this difference reflects the impact of random measurement error, systematic differences in the sample of participants, or the changes that were made to the statement pool. In terms of the evaluative-implications model, a stepping stone for future research might thus be to identify conditions that for a negative and a positive context reliably lead to equivalent (item and category) memory parameters.

Let us conclude with a silver lining: From a meta-perspective, the present manuscript reinforces our argument for the necessity of studies like our Study 3. Prior to trying to understand the mechanisms that may have caused an effect, it is important for the advancement of psychological science to assess its robustness. The replication attempt in Study 3 brought to light that we need to understand the conditions under which a contextual modulation of the link between categorization and prejudice appears reliably first before any investigation into possible underlying theoretical mechanisms makes sense. As explained above, theoretically, a contextual modulation of the categorization-evaluation link is quite plausible: Both categorization and prejudice are heavily impacted by contextual variables, and what participants attend to in a given moment strongly influences their evaluations. But as long as future researchers do not find an experimental setup in which this can be shown reliably, we need to acknowledge that we do not have direct evidence for this hypothesis.

8. Conclusion

The present manuscript revisited the link between categorization and prejudice. Based on the correlational pattern found in Study 1, we developed the hypothesis that the evaluative implications of the context in which people are encountered might shape how categorization and social evaluations are intertwined. The data from Study 2 were well in line with this hypothesis, but the data from Study 3 with a larger sample were not. Changes to the WSW procedure meant to improve on its validity now make it somewhat difficult to attribute this failure to replicate the previous results. Until these issues are resolved, there is little evidence for the claim that the evaluative context meaningfully impacts the link between categorization and evaluation. Thus, Park and Judd's (2005) conclusion that there is little evidence for a robust graded link between the strength of social categorization and evaluation still stands.

Open science practices

The Stage-1 version of this registered report can be found at https://osf.io/preprints/psyarxiv/rjnpk_v1, also see https://aspredicted.org/tmdm-sz26.pdf for a summary of our preregistered analyses for Study 3. As mentioned above, the Online Repository at https://osf.io/dk3zp/?view_only=7bc33ca589d045fa809ed811b38ddee1 includes the computer programs and other materials presented to participants in the course of the study, as well as the raw data. For all studies reported in this manuscript, we report all measures, manipulations, and exclusions in these studies. Sample sizes were determined before any data analysis.

[6] Results are similar for the sample that included the BIPoC participants that fulfilled all other inclusion criteria, for the sample that included the excess participants not included in the main analysis, and for the replication of the main analysis that relies on the first half of EDT trials only. In particular, in all three analyses, there is a significant priming effect only in the disadvantaged-Blacks condition, and the estimate for racial category memory is larger for the disadvantaged-Blacks condition ($0.30 < d_{race} < 0.31$) than for the advantaged-Blacks condition ($0.17 < d_{race} < 0.19$). Most importantly, there is never a positive regression coefficient in the disadvantaged-Blacks condition, never a negative regression coefficient in the advantaged-Blacks condition, and the 95 % HDIs for their differences in the joint analysis always include 0.

CRediT authorship contribution statement

Manuel Becker: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sarah Teige-Mocigemba:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jeffrey W. Sherman:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition, Conceptualization. **Karl Christoph Klauer:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Funding

The research reported in this paper was supported by the Anneliese Maier Research Award of the Alexander von Humboldt Foundation to Jeffrey W. Sherman.

Declaration of competing interest

The link between social categorization and spontaneous social evaluations: A matter of the evaluative implications of the situation?

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2025.104842.

References

Alinor, M., & Tinkler, J. (2021). TAKE OFF YOUR HOODIE: Assessing how professional attire influences the perception of black men as threatening. *Du Bois Review: Social Science Research on Race, 18*, 97–117. https://doi.org/10.1017/S1742058X21000072

Allport, G. W. (1954). *The nature of prejudice*. Cambridge, Mass: Addison-Wesley Pub. Co.

Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality and Social Psychology Bulletin, 38*, 1194–1208. https://doi.org/10.1177/0146167212446835

Barden, J., Maddux, W. W., Petty, R. E., & Brewer, M. B. (2004). Contextual moderation of racial bias: The impact of social roles on controlled and automatically activated attitudes. *Journal of Personality and Social Psychology, 87*, 5–22. https://doi.org/10.1037/0022-3514.87.1.5

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin & Review, 6*, 57–86. https://doi.org/10.3758/BF03210812

Bodenhausen, G. V., Kang, S. K., & Peery, D. (2012). Social categorization and the perception of social groups. In S. Fiske, & C. N. Macrae (Eds.), *The SAGE handbook of social cognition* (pp. 311–329). SAGE Publications Ltd.

Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association, 82*, 106–111. https://doi.org/10.2307/2289130

Casper, C., Rothermund, K., & Wentura, D. (2010). Automatic stereotype activation is context dependent. *Social Psychology, 41*, 131–136. https://doi.org/10.1027/1864-9335/a000019

Castano, E., Yzerbyt, V., Bourguignon, D., & Seron, E. (2002). Who may enter? The impact of in-group identification on in-group/out-group categorization. *Journal of Experimental Social Psychology, 38*, 315–322. https://doi.org/10.1006/jesp.2001.1512

Clark-Carter, D. (2004). *Quantitative psychological research: A student's handbook*. New York: Psychology Press.

Correll, J., Guillermo, S., & Vogt, J. (2014). On the flexibility of attention to race. *Journal of Experimental Social Psychology, 55*, 74–79. https://doi.org/10.1016/j.jesp.2014.05.013

Czopp, A. M., & Monteith, M. J. (2006). Thinking well of African Americans: Measuring complimentary stereotypes and negative prejudice. *Basic and Applied Social Psychology, 27*, 233–250. https://doi.org/10.1207/s15324834basp2803_3

De Santis, F., & Gubbiotti, S. (2021). Sample size requirements for calibrated approximate credible intervals for proportions in clinical trials. *International Journal of Environmental Research and Public Health, 18*, 595, 10.3390%2Fijerph18020595.

Deffenbacher, D. M., Park, B., Judd, C. M., & Correll, J. (2009). Category boundaries can be accentuated without increasing intergroup bias. *Group Processes & Intergroup Relations, 12*, 175–193. https://doi.org/10.1177/1368430208101055

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

Eagly, & Diekman. (2005). What is the problem? Prejudice as an attitude-in-context. In J. F. Dovidio, P. Glick, & L. Rudman (Eds.), *On the nature of prejudice: Fifty years after Allport* (pp. 19–35). Malden, MA: Blackwell Publishing. https://doi.org/10.1002/9780470773963.ch2.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: race, crime, and visual processing. *Journal of Personality and Social Psychology, 87*, 876–893. https://doi.org/10.1037/0022-3514.87.6.876

Ege, M., & Hurley, A. W. (2015). Introduction: Special issue on African-Americanophilia in Germany. *PORTAL Journal of Multidisciplinary International Studies, 12*. https://doi.org/10.5130/portal.v12i2.4358

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

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes. *Journal of Personality and Social Psychology, 69*, 1013–1027. https://doi.org/10.1037/0022-3514.69.6.1013

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*, 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology, 23*, 1–74. https://doi.org/10.1016/S0065-2601(08)60317-2

Flade, F., Klar, Y., & Imhoff, R. (2019). Unite against: A common threat invokes spontaneous decategorization between social categories. *Journal of Experimental Social Psychology, 85*, Article 103890. https://doi.org/10.1016/j.jesp.2019.103890

Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., & Ambady, N. (2011). Looking the part: Social status cues shape race perception. *PLoS One, 6*, Article e25107. https://doi.org/10.1371/journal.pone.0025107

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., & Bodenhausen, G. V. (2017). Beyond persons and situations: An interactionist approach to understanding implicit Bias. *Psychological Inquiry, 28*, 268–272. https://doi.org/10.1080/1047840X.2017.1373546

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 283–310). New York, NY: Cambridge University Press.

Gawronski, B., & Sritharan, R. (2010). Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 216–240). New York, NY: Guildford Press.

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*, 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: An integrative account of the other-race recognition deficit. *Psychological Review, 117*, 1168–1187. https://doi.org/10.1037/a0020463

Hughes, S., Cummins, J., & Hussey, I. (2021). Effects on the affect misattribution procedure are strongly moderated by influence awareness. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01879-4

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology, 41*, 137–146. https://doi.org/10.1027/1864-9335/a000020

Judd, C., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54–69. https://doi.org/10.1037/a0028347

Kay, A. C., Day, M. V., Zanna, M. P., & Nussbaum, A. D. (2013). The insidious (and ironic) effects of positive stereotypes. *Journal of Experimental Social Psychology, 49*, 287–291.

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika, 75*, 70–98. https://doi.org/10.1007/s11336-009-9141-0

Klauer, K. C., & Becker, M. (2024). Latent state-trait analyses for process models of implicit measures. In J. A. Krosnick, T. H. Stark, & A. L. Scott (Eds.), *The Cambridge handbook of implicit Bias and racism*. Cambridge, England: Cambridge University Press.

Klauer, K. C., Ehrenberg, K., & Wegener, I. (2003). Crossed categorization and stereotyping: Structural analyses, effect patterns, and dissociative effects of context relevance. *Journal of Experimental Social Psychology, 39*, 332–354. https://doi.org/10.1016/S0022-1031(03)00017-9

Klauer, K. C., Hölzenbein, F., Calanchini, J., & Sherman, J. W. (2014). How malleable is categorization by race? Evidence for competitive category use in social categorization. *Journal of Personality and Social Psychology, 107*, 21–40. https://doi.org/10.1037/a0036609

Klauer, K. C., Teige-Mocigemba, S., & Spruyt, A. (2009). Contrast effects in spontaneous evaluations: a psychophysical account. *Journal of Personality and Social Psychology, 96*, 265–287. https://doi.org/10.1037/a0013248

Klauer, K. C., & Wegener, I. (1998). Unraveling social categorization in the "who said what?" paradigm. *Journal of Personality and Social Psychology, 75*, 1155–1178. https://doi.org/10.1037/0022-3514.75.5.1155

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis*. Burlington, MA: Academic Press.

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

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

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian data analysis for cognitive science: A practical course*. New York: Cambridge University Press.

Levine, M., Prosser, A., & Evans, D. (2005). Identity and emergency intervention: How social group membership and inclusiveness of group boundaries shapes helping behavior. *Personality and Social Psychology Bulletin, 31*, 443–453. https://doi.org/10.1177/0146167204271651

Maddux, W. W., Barden, J., Brewer, M. B., & Petty, R. E. (2005). Saying no to negativity: The effects of context and motivation to control prejudice on automatic evaluative responses. *Journal of Experimental Social Psychology, 41*, 19–35. https://doi.org/10.1016/j.jesp.2004.05.002

Madon, S., Guyll, M., Aboufadel, K., Montiel, E., Smith, A., Palumbo, P., & Lee, J. (2001). Ethnic and National Stereotypes: The Princeton trilogy revisited and revised. *Personality and Social Psychology Bulletin, 27*, 996–1010. https://doi.org/10.1177/0146167201278007

Mann, T. C., Cone, J., Heggeseth, B., & Ferguson, M. J. (2019). Updating implicit impressions: New evidence on intentionality and the affect misattribution procedure. *Journal of Personality and Social Psychology, 116*, 349–374. https://doi.org/10.1037/pspa0000146

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika, 80*, 205–235. https://doi.org/10.1007/s11336-013-9374-9

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General, 132*, 455–469. https://doi.org/10.1037/0096-3445.132.3.455

Monroe, B. M., Koenig, B. L., Wan, K. S., Laine, T., Gupta, S., & Ortony, A. (2018). Re-examining dominance of categories in impression formation: A test of dual-process models. *Journal of Personality and Social Psychology, 115*, 1–30. https://doi.org/10.1037/pspa0000119

Park, B., & Judd, C. (2005). Rethinking the link between categorization and prejudice within the social cognition perspective. *Personality and Social Psychology Review, 9*, 108–130. https://doi.org/10.1207/s15327957pspr0902_2

Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention invention and the affect misattribution procedure: Reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin, 39*, 375–386. https://doi.org/10.1177/0146167212475225

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*, 277–293. https://doi.org/10.1037/0022-3514.89.3.277

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

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8–13. https://doi.org/10.1016/J.JNEUMETH.2006.11.017

Phills, C. E., Hahn, A., & Gawronski, B. (2020). The bidirectional causal relation between implicit stereotypes and implicit prejudice. *Personality and Social Psychology Bulletin, 46*, 1318–1330. https://doi.org/10.1177/0146167219899234

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*, 573–604. https://doi.org/10.3758/BF03196750

Schaller, M., Park, J. H., & Mueller, A. (2003). Fear of the dark: Interactive effects of beliefs about danger and ambient darkness on ethnic stereotypes. *Personality and Social Psychology Bulletin, 29*, 637–649. https://doi.org/10.1177/0146167203029005008

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*, 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition, 25*, 638–656. https://doi.org/10.1521/soco.2007.25.5.638

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559–569. https://doi.org/10.1177/0956797614567341

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology, 1*, 149–178. https://doi.org/10.1002/ejsp.2420010202

Taylor, S. E., Fiske, S. T., Etcoff, N. J., & Ruderman, A. J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology, 36*, 778–793. https://doi.org/10.1037/0022-3514.36.7.778

Teige-Mocigemba, S., Becker, M., Sherman, J. W., Reichardt, R., & Klauer, K. C. (2017). The affect misattribution procedure: In search of prejudice effects. *Experimental Psychology, 64*, 215–230. https://doi.org/10.1027/1618-3169/a000364

Todd, A. R., Johnson, D. J., Lassetter, B., Neel, R., Simpson, A. J., & Cesario, J. (2021). Category salience and racial bias in weapon identification: A diffusion modeling approach. *Journal of Personality and Social Psychology, 120*, 672–693. https://doi.org/10.1037/pspi0000279

Vescio, T. K., Judd, C. M., & Kwan, V. S. Y. (2004). The crossed categorization hypothesis: Evidence of reductions in the strength of categorization, but not intergroup bias. *Journal of Experimental Social Psychology, 40*, 478–496. https://doi.org/10.1016/j.jesp.2003.09.005

Voss, A., Leonhart, R., & Stahl, C. (2007). How to make your own response boxes: A step-by-step guide for the construction of reliable and inexpensive parallel-port response pads from computer mice. *Behaviour Research Methods, 39*, 797–801. https://doi.org/10.3758/BF03192971

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