Stimulus Confounds in Implicit and Explicit Measures of Racial Bias

Aline da Silva Frost
University of California, Davis

Alison Ledgerwood
University of California, Davis

Paul W. Eastwick
University of California, Davis

Bertram Gawronski
University of Texas at Austin

Implicit measures often show dissociations from explicit measures, including low correlations, distinct antecedents, and distinct behavioral correlates. Interpretations of these dissociations referring to measurement types presuppose that the distinction between implicit and explicit measures is not confounded with other stimulus-related differences. However, in research on racial bias, explicit measures often use verbal category labels, whereas implicit measures include images of specific exemplars. The current work addressed this confound by investigating associations between implicit and explicit measures of racial bias that include verbal category labels and images of exemplars, respectively. Experiments 1 and 2 tested whether implicit and explicit measures show stronger associations when they correspond in terms of their stimuli. Experiments 3 and 4 investigated whether previously obtained moderators of implicit-explicit relations qualify the association between measures that focus on different types of stimuli, rather than implicit and explicit measures *per se.* While the overall results are mixed, our analysis suggests that more attention should be paid to stimulus confounds when studying dissociations between implicit and explicit measures.

Keywords: categories; exemplars; implicit measures; prejudice; racial bias

Arguably, one of the most impactful inventions within the field of psychology during the last three decades has been the development of implicit measures (for reviews, see Gawronski et al., 2020; Greenwald & Lai, 2020). A central feature of implicit measures is that they allow researchers to infer evaluative responses to attitudinal stimuli from objective performance indicators (e.g., speed, accuracy) rather than direct self-reports. Implicit measures have gained popularity partly because they often show dissociations from explicit self-report measures, including low correlations between the two kinds of measures (Nosek, 2005), distinct antecedents (Gawronski & Bodenhausen, 2006), and distinct behavioral correlates (Friese et al., 2008). In research on prejudice and stereotyping, these dissociations are often interpreted as evidence for the idea that implicit measures capture social biases that people are unwilling or unable to report (Greenwald & Banaji, 1995; Fazio et al., 1995).

Although claims that implicit measures provide a window into unconscious biases are controversial (see Corneille & Hütter, 2020; Fazio & Olson, 2003; Gawronski et al., 2022a; Hahn et al., 2014), many researchers agree that responses on implicit measures tend to be unintentional and difficult to control (e.g., De Houwer & Boddez, 2022; Gawronski et al., 2022b; Melnikoff & Kurdi, 2022; Olson & Gill, 2022; Ratliff & Smith, 2022). Thus, researchers have assumed that dissociations between responses on implicit measures (presumably unintentional and difficult to control) and responses on explicit measures (presumably intentional and easy to control) can be interpreted in terms of the two aspects of automaticity: intentionality and controllability (Bargh, 1994; Moors & De Houwer, 2006). Yet, such an interpretation presupposes that the distinction between implicit and explicit measures is not confounded with other important differences between measures.

In the current work, we identify an important confound between many implicit and explicit measures of racial bias: the confound between type of measure and type of stimuli. Whereas explicit measures of racial bias typically use verbal category labels, most implicit measures include images of specific exemplars (e.g., Fazio et al., 1995; Greenwald et al., 1998; Payne et al., 2005; for a notable exception, see Wittenbrink et al., 1997). Pulling apart and taking seriously confounds between type of measure (i.e., implicit vs. explicit) and type of stimuli (i.e., verbal category labels vs. images of specific exemplars) is essential for gaining insights into the causes and consequences of racial bias, with important implications for both theory and practice.

Measurement and Construct

Three implicit measures stand out in terms of the frequency with which they have been used in research on racial bias (Greenwald & Lai, 2020; Nosek et al., 2011): the Evaluative Priming Task (EPT; Fazio et al., 1995), the Implicit Association Test (IAT; Greenwald et al., 1998), and the Affect Misattribution Procedure

(AMP; Payne et al., 2005). In a typical EPT, participants are briefly presented with an attitudinal prime stimulus (e.g., an image of a White or Black face), which is followed by a positive or negative target word. Participants' task is to indicate as quickly as possible whether the target word is positive or negative. The idea underlying the EPT is that quick and accurate responses to the target words should be facilitated when they are evaluatively congruent with participants' attitude toward the prime stimulus. In contrast, quick and accurate responses to the target words should be impaired when they are evaluatively incongruent with participants' attitude toward the prime stimulus (Fazio, 2001). For example, if a person holds more favorable attitudes toward White people than Black people, this person should be faster and more accurate in identifying the valence of positive words when the person has been primed with an image of a White person compared to when they have been primed with an image of a Black person. Conversely, the person should be slower and less accurate in identifying the valence of negative words when they have been primed with an image of a White person compared to when they have been primed with an image of a Black person.

The most prominent implicit measure in research on racial bias is the IAT (Greenwald et al., 1998). In the critical blocks of the IAT, participants are asked to complete two binary categorization tasks that are combined in a manner that is either congruent or incongruent with the to-be-measured attitude. For example, in the commonly used race IAT, participants may be asked to categorize images of Black and White faces in terms of their race and positive and negative words in terms of their valence. In one critical block of the task, participants are asked to press one response key for Black faces and negative words, and another response key for White faces and positive words (i.e., prejudice-congruent block). In the other critical block, participants are asked to complete the same categorization tasks with a reversed key assignment for the faces, such that they have to press one response key for White faces and negative words, and the other response key for Black faces and positive words (i.e., prejudice-incongruent block). The basic idea underlying the IAT is that responses in the task should be facilitated when two mentally associated concepts are mapped onto the same response key. For example, a person who has more favorable attitudes toward White people than Black people should show faster and more accurate responses when White faces share the same response key with positive words and when Black faces share the same response key with negative words, compared with the reversed mapping.

The AMP was designed to combine the structural advantages of the EPT with the superior psychometric properties of the IAT (Payne et al., 2005). Two central differences of the AMP are that (1) the target stimuli in the AMP are ambiguous and (2) participants are asked to report their subjective evaluations of the targets. The basic idea is that participants may misattribute the affective feelings elicited by the prime stimuli to the neutral targets, and therefore judge the targets more favorably when they were primed with a positive stimulus than when they were primed with a negative stimulus (for a review, see Payne & Lundberg, 2014). For example, in an AMP to measure racial attitudes, participants may be asked to indicate whether they find Chinese ideographs visually more pleasant or visually less pleasant than average after being primed with images of Black versus White faces. A preference for White over Black people would be indicated by a tendency to evaluate the Chinese ideographs more favorably when the ideographs followed the presentation of a White face than when they followed the presentation of a Black face.

The development of implicit measures paved the way for considerable theory and research, which can be characterized by three core themes (for a review, see Gawronski et al., 2020). First, research on the relation between implicit and explicit measures revealed that correlations between the two are often quite low (for meta-analyses, see Cameron et al., 2012; Hofmann, Gawronski, et al., 2005). Second, research on the antecedents of responses on implicit and explicit measures suggests that they may differ in their sensitivity to different kinds of information (for a review, see Gawronski & Bodenhausen, 2006). Finally, research on behavioral correlates of implicit and explicit measures suggests that they may predict different kinds of behaviors (e.g., unintentional vs. intentional behavior) and behavior under different contextual conditions (e.g., high vs. low cognitive load; for a review, see Friese et al., 2008).

The available evidence for low correlations between implicit and explicit measures, distinct antecedents, and distinct behavioral correlates have led many researchers to conclude that implicit and explicit measures capture related but distinct constructs (e.g., Greenwald & Nosek, 2009; Nosek & Smyth, 2007). Such a conclusion would seem justified to the extent that the relevant evidence was obtained with implicit and explicit measures that have conceptually equivalent content (e.g., use the same stimuli) aside from the implicit and explicit nature of their measurement approaches. However, a closer inspection of the literature suggests that this is not always the case, with research on racial bias standing out as a particularly problematic area (for discussions, see Axt, 2018; Cooley & Payne, 2017; Gawronski, 2019; Payne et al., 2008). To the extent that an implicit measure has little or no content correspondence with an explicit measure, their relation can be expected to be low for important but often overlooked methodological reasons (Ajzen & Fishbein, 1977). In such cases, it would be premature to interpret their weak relation as evidence for the hypothesis that implicit and explicit measures capture distinct constructs (e.g., Greenwald & Nosek, 2009; Nosek & Smyth, 2007). Similarly, if type of measure is confounded with different contents, any finding suggesting distinct antecedents or distinct behavioral correlates remains ambiguous, because the obtained dissociation could be due to either (1) the implicit versus explicit nature of the measures or (2) the different contents of the two measures.

Verbal Category Labels and Images of Specific Exemplars

In our view, one of the most important confounds between implicit and explicit measures of racial bias in the existing literature is that they typically have differed in terms of the stimuli presented to participants (see Cooley & Payne, 2017; Gawronski, 2019; Irving & Smith, 2020). A common practice in racial bias research using implicit and explicit measures is to use images of specific exemplars as attitudinal stimuli in the implicit measure (e.g., images of Black and White faces in the EPT, the IAT, and the AMP) and to use verbal category labels in the explicit measure (e.g., the category labels *Black people* and *White people* in feeling thermometer ratings or semantic differentials). Although it seems possible that responses to verbal category labels are systematically related to those evoked by images of specific exemplars of a given category, the two kinds of stimuli differ markedly in terms of their levels of abstraction.

Research guided by construal level theory suggests that both categories and words tend to prompt more abstract mental representations of an object, whereas exemplars and images tend to prompt more concrete mental representations of an object (Fujita et al., 2008; Henderson, 2013; Rim et al., 2015). Furthermore, thinking about objects at different levels of abstraction can have different consequences for a range of cognitive processes and evaluative judgments (Fujita & Han, 2009; Ledgerwood et al., 2010, 2019; Soderberg et al., 2015; Wakslak & Trope, 2009). For example, evaluations of exemplar images may reflect people's experienced responses to concrete targets, whereas evaluations of category labels may reflect people's abstract ideas about their likes and dislikes (Eastwick et al., 2019; Ledgerwood et al., 2018). In fact, evaluations at concrete versus abstract levels often show low correlations, have different antecedents, and predict different downstream consequences—even if all the measures derive from explicit self-reports (da Silva Frost et al., 2024; Ledgerwood & Wang, 2018).[1]

The Current Research

In sum, previous research has concluded that implicit and explicit measures capture related but distinct constructs based on evidence that implicit and explicit measures have low correlations, distinct antecedents, and distinct behavioral correlates. However, our reasoning above suggests that such conclusions are premature in the absence of more compelling evidence that the observed dissociations are indeed rooted in differences between implicit and explicit measures rather than differences in the types of stimuli presented in the two kinds of measures. If research disentangling the confound between measurement type and measurement content suggests a significant role of measurement content that is independent of the distinction between implicit and explicit measures, the obtained evidence would necessitate a reassessment of a substantial body of earlier findings.

In the current work, we addressed these issues for racial-bias applications of the AMP and the IAT. Our focus on the two measures was based on three considerations. First, they are the two implicit measures with the highest internal consistencies (Gawronski & De Houwer, 2014; Greenwald & Lai, 2020), which is methodologically imperative for the correlational designs of the current studies (see Koppehele-Gossel et al., 2020). Second, the two measures are based on different underlying mechanisms (Gawronski & De Houwer, 2014; Nosek et al., 2011), which is essential for testing the generality or specificity of our results (see Gawronski et al., 2008). Third, the two measures have unique strengths and weaknesses, allowing us to complement the weaknesses of one measure with the strengths of the other.

The main goal of Experiments 1 and 2 was to test whether implicit and explicit measures of racial bias show stronger associations when they correspond (versus do not correspond) in terms of their stimuli (i.e., verbal category labels vs. images of specific exemplars). Toward this end, Experiments 1 and 2 assessed the correlations between implicit and explicit

[1] Like Ajzen and Fishbein (1977, 2005), we argue that evaluative responses may seem to have weak correspondence in cases where researchers specify the object of evaluation in mismatching ways. For example, according to Ajzen and Fishbein, a general attitude (e.g., attitudes toward environmentalism) might fail to predict a specific behavior (e.g., voting on a city ordinance that would require composting), because the two attitude objects are not specified at the same level (see also Ledgerwood et al., 2010). It is worth noting, however, that the distinction we make here between evaluations of verbal category labels (e.g., a person's evaluation of the category *Black people*) and evaluations of images of specific exemplars (e.g., a person's average evaluation of a set of Black faces) is different in that, in Ajzen and Fishbein's (1977) work, these two measures of evaluations were actually treated as interchangeable methods for assessing the same general attitude construct (see Ledgerwood et al., 2018).

measures of racial bias that either match or mismatch in terms of their contents. In Experiment 1, participants completed an AMP measure of either (1) responses to images of Black and White faces (i.e., implicit exemplar-image measure) or (2) responses to verbal category labels like *Black people* and *White people* (i.e., implicit category-label measure). In addition, participants completed a feeling thermometer measure of either (1) responses to images of Black and White faces (i.e., explicit exemplar-image measure) or (2) responses to verbal category labels (i.e., explicit category-label measure). We tested whether implicit-explicit relations are stronger when the measures correspond in terms of their content compared to when they do not.

In Experiment 2, participants completed a typical race IAT, a feeling thermometer with images of specific exemplars, and a feeling thermometer with verbal category labels. The IAT is different from many other implicit measures, in that it includes images of specific faces as target stimuli and verbal category labels for the response options. Thus, variance in the IAT might be jointly driven by responses to images of specific exemplars and responses to verbal category labels. Based on these considerations, we tested whether the association between IAT scores and each of the two explicit measures remains significant when controlling for the respective other explicit measure. A significant association between IAT scores and an explicit category-label measure, controlling for an explicit exemplar-image measure, would suggest that responses to verbal category labels play a unique role that is not accounted for by responses to images of specific exemplars. Conversely, a significant association between IAT scores and an explicit exemplar-image measure, controlling for an explicit category-label measure, would suggest that responses to images of specific exemplars play a unique role that is not accounted for by responses to verbal category labels.

Expanding on the findings of the first two studies, Experiments 3 and 4 investigated whether presumed moderators of implicit-explicit relations qualify the correspondence between measures with different contents, rather than implicit and explicit measures *per se*. Past research has found that individuals who report a strong motivation to control prejudiced reactions show weaker relations between implicit and explicit measures of bias, compared to individuals who report a weak motivation to control prejudiced reactions (e.g., Degner & Wentura, 2008; Dunton & Fazio, 1997; Fazio et al., 1995; Gawronski et al., 2003; Payne et al., 2005). In Experiments 3 and 4, we tested the alternative hypothesis that this pattern may be driven by an attenuating effect of motivation to control prejudice on the correspondence between responses to verbal category labels and responses to images of specific exemplars, independent of whether responses to the two kinds of stimuli are captured with an implicit or an explicit measure.

To this end, participants in Experiments 3 and 4 completed Dunton and Fazio's (1997) Motivation to Control Prejudice Reactions Scale (MCPRS) and a typical feeling thermometer with category labels (i.e., explicit category-label measure). In Experiment 3, half of the participants additionally completed a typical AMP with Black and White faces (i.e., implicit exemplar-image measure), while the remaining half completed an feeling thermometer with the same Black and White faces (i.e., explicit exemplar-image measure). We tested the hypothesis that, regardless of whether the exemplar-image measure is implicit or explicit, the relation between the exemplar-image measure and the explicit category-label measure decreases as MCPRS scores increase. Participants in Experiment 4 completed a typical IAT with images of faces as target stimuli and verbal category labels for the response options, a typical feeling thermometer with category labels (i.e., explicit category-label measure), and a feeling thermometer with Black and White faces (i.e., explicit exemplar-image measure). We tested the prediction that MCPRS scores moderate not only the relation between racial bias on the IAT and the explicit category-label measure (replicating earlier findings), but also the relation between the explicit exemplar-image measure and the explicit category-label measure. To the extent that our predictions in Experiments 3 and 4 are confirmed, the findings would suggest that motivation to control prejudiced reactions moderates relation between responses to different types of stimuli (i.e., images of specific exemplars vs. verbal category labels) rather than different types of measurement instruments (i.e., implicit vs. explicit).

Open Practices

For all studies, we report how we determined our sample sizes, all data exclusions, all manipulations, and all measures. The data, analysis codes, and research materials for all studies are available at https://osf.io/d9q3g/. For all studies, we preregistered our analysis plan, including target sample size and exclusion criteria. Hyperlinks to the preregistrations are reported in the Methods sections of each study.

Experiment 1

The main goal of Experiment 1 was to test if the correlations between explicit and implicit measures of racial bias are higher when they correspond (versus do not correspond) in terms of their stimuli (images of specific exemplars vs. verbal category labels). To this end, each participant completed one AMP and one feeling-thermometer measure, either of which could include verbal category labels or images of specific exemplars. If correspondence in terms of stimuli is indeed relevant for interpreting relations between the AMP and explicit measures, correlations between measures should be higher when participants completed two measures with matching stimuli (see A and C in Figure 1) than when they completed two measures with mismatching stimuli (see B and D in Figure 1).

Method

Preregistration

The analysis plan was preregistered at https://aspredicted.org/ZUC_ABW.

Participants

We preregistered a target sample size of 1250, so that after an estimated 50 exclusions, we would have 300 participants per condition. We based this target sample on an a priori power analysis in G*Power (Faul et al., 2007), which indicated that we would need 300 participants per condition in order to achieve 80% power to detect a difference between correlations of *q* = .23 between any two of the four conditions. Participants were MTurk workers over the age of 18 who completed the experiment on Inquisit. The raw dataset has 1252 rows. Out of these, 1 is missing data on key measures, 2 are incompletes, and 4 are duplicates. We preregistered that we would exclude participants who either (1) reported knowing the meaning of the ideographs in the AMP (answering *yes* to a *yes*/*no* question), or (2) provided a nonsensical response to a Winograd-like attention check, or (3) pressed the same key on all trials of the AMP. We excluded 124 participants who met one or more of the exclusion criteria (*n* = 83 reported knowing the ideographs, *n* = 33 failed the Winograd-like attention check, *n* = 38 pressed the same key; of these, 30 participants met more than one criteria). Analyses were conducted on the remaining 1121 participants (624 women, 486 men, and 11 people who chose a different option; M_{age} = 39.3, *SD* = 12.8; 80.8% White, 8% Black, 8.1% Hispanic or Latino/a/e, 3.5% Black and White Biracial, 2.1% South Asian, 3.7% East Asian, and 1.7% a different identity). For the smallest cell sizes in the four experimental conditions, the final sample provides 80% power to detect a difference between correlations of *q* = .25.

Procedure and Materials

Participants were randomly assigned to one of four between-subjects conditions. In each condition, participants completed one explicit and one implicit measure (the order of the two measures was counterbalanced). The four conditions were: (1) explicit exemplar-image and implicit exemplar-image; (2) explicit exemplar-image and implicit category-label; (3) explicit category-label and implicit category-label; and (4) explicit category-label and implicit exemplar-image (see Figure 1). For the category-label measures, we used the labels *African American*, *European American*, *Black people*, and *White people*. For the exemplar-image measures, we used 20 faces of Black and White women and men (5 Black women, 5 White women, 5 Black men, and 5 White men; all images taken from Minear & Park, 2004). Afterwards, participants completed an attention check and demographic questions.[2]

Explicit Exemplar-Image Measure. The explicit exemplar-image measure was a feeling thermometer in which participants rated their feelings toward each face on a 7-point scale ranging from 1 (*very negative*) to 7 (*very positive*). The faces were presented one at a time in random order. To create an index of explicit preference for White over Black exemplars, we computed the average explicit ratings of the ten White faces and the average explicit ratings of the ten Black faces, and then subtracted the average ratings of Black faces from the average ratings of White faces.

Explicit Category-Label Measure. The explicit category-label measure consisted of six feeling thermometer ratings. On four of the six items, participants rated their feelings toward each of the four categories (i.e., African American, European American, Black people, and White people) on 5-point scales ranging from 1 (*very negative*) to 5 (*very positive*). On the remaining two items, participants rated their relative preference for one category over the other (i.e., African American vs. European American; Black people vs. White people) on 9-point scales ranging -4 (*strongly prefer African Americans / Black people over European Americans / White people*) to +4 (*strongly prefer European Americans / White people over African Americans / Black people*). To create an index of explicit preference for White over Black categories, we first computed difference scores for each of the two pairs of single-category feeling thermometers, and then averaged the resulting two difference scores with the two relative preference ratings (all coded in the same direction).

Implicit Exemplar-Image Measure. The implicit exemplar-image measure was an AMP measuring responses to the same faces used in the explicit exemplar-image measure described above. On each trial of the task, participants were first presented with a fixation cross for 500 ms, which was replaced by an image of either a Black or a White face for 75 ms. The presentation of the face prime was followed by a blank

[2] In all studies, we also included a funnel debriefing at the end to probe for suspicion, but this measure was not part of the preregistered exclusion criteria.

screen for 125 ms, after which a Chinese ideograph appeared for 100 ms. The Chinese ideograph was then replaced by a black-and-white pattern mask, and participants had to indicate whether they considered the Chinese ideograph as more pleasant or less pleasant than the average Chinese ideograph. The pattern mask remained on the screen until participants gave their response. Participants were asked to press a right-hand key (*I*) if they considered the Chinese ideograph as more pleasant than average, and a left-hand key (*E*) if they considered the Chinese ideograph as less pleasant than average. Following the instructions employed by Payne et al. (2005), participants were told that the faces can sometimes bias people's responses to the Chinese ideographs, and that they should try their absolute best not to let the faces influence their judgments of the Chinese ideographs. Each of the 20 face primes was presented twice, summing up to a total of 40 trials. As target stimuli, we used 40 distinct Chinese ideographs from Payne et al. (2005). Order of trials was randomized for each participant. To create an index of implicit preference for White over Black exemplars, we calculated the proportion of *pleasant* responses to target stimuli for each prime type (i.e., Black face vs. White face), and then subtracted the proportion score for Black face primes from the proportion score for White face primes.

Implicit Category-Label Measure. The implicit category-label measure was an AMP measuring responses to the same four category labels used in the explicit category-label measure described above. The procedural details were identical to the AMP measuring responses to exemplar-images, the only difference being that we used the four category labels as primes (i.e., *African American*, *European American*, *Black people*, and *White people*) instead of Black and White faces. Each of the category labels was presented ten times across 40 trials. To create an index of implicit preference for White over Black categories, we calculated the proportion of *pleasant* responses to target stimuli for each prime type, and then subtracted the proportion score for Black category primes from the proportion score for White category primes.

Knowledge About the Meaning of the Ideographs. In line with previous AMP practices (e.g., Gawronski & Ye, 2014), we asked participants *Do you know the meaning of the Chinese ideographs we showed you in the concentration task?* (Yes/No) in the demographics section at the end of the study.

Winograd-like Attention Check. To identify bots and inattentive participants, we used a Winograd-like attention check, which is part of the standard operating procedures in the first and second authors' lab. This check involves text interpretation based on the structure of a Winograd schema (used to assess human-like reasoning; Levesque et al., 2012). To this end, participants read the following story: *An elderly man had the dream of watching the American female soccer team playing for their country. His grandson bought him a ticket to travel to Brazil during the 2016 Olympics as a gift. When he woke up on the day of his birthday and received the ticket, he cried of happiness. Thinking about the game, he hoped to be able to see legends such as Carli Lloyd, Megan Rapinoe, and Marta up close.* Next, they answered two open-ended questions about the story (*Who got a birthday gift?* and *What does the man expect for Summer 2016?*). Participants were excluded if they gave nonsensical answers (e.g., "unhappily") to either question, as coded by a researcher without knowledge of the results.

Results

Descriptive statistics of the four measures are presented in Table 1. If measurement content affects correspondence, correlations between measures with matching stimuli (A and C in Figure 1) should be higher than the correlations between measures with mismatching stimuli (B and D in Figure 1). Following our preregistered analysis plan, we conducted z-score tests to examine if there were significant differences between the four pairs of correlations (see Table 2). As predicted, the matching content correlation between the two category-label measures was significantly higher than the mismatching content correlation between the implicit category-label measure and the explicit exemplar-image measure (q = .20, z = 2.37, p = .018). Although the pattern of differences were in the expected direction for the other three comparisons (i.e., matching content > mismatching content), we did not find a significant difference when comparing the matching content correlation between the two exemplar-image measures with the mismatching content correlation between the implicit category-label measure and the explicit exemplar-image measure (q = .11, z = 1.22, p = .223), when comparing the matching content correlation between the two category-label measures with the mismatching content correlation between the implicit exemplar-image measure and the explicit category-label measure (q = .15, z = 1.89, p = .058), and when comparing the matching content correlation between the two exemplar-image measures with the mismatching content correlation between the implicit exemplar-image measure and the explicit category-label measure (q = .07, z = 0.71, p = .478).

Discussion

In this experiment, we tested the hypothesis that explicit and implicit measures of racial bias show higher correlations when the measures correspond (versus do not correspond) in terms of their stimuli. Although the pattern of differences was in the expected direction for all four comparisons, the differences were fairly small and only one of the four comparisons reached statistical significance. Of course, it is possible that some of the comparisons did not reach statistical significance due to insufficient statistical power. The effect sizes for the four preregistered comparisons were $q = .20$, $q = .11$, $q = .15$, and $q = .07$, which average to be $q = .13$ (i.e., a small effect size, Cohen, 1992). Detecting an effect of this size would require $N = 1{,}836$ to obtain 80% power in a two-tailed test with an α-level of .05. Thus, it seems as though the impact of stimulus confounds is not likely to be zero in the AMP, although the effect seems to be relatively small and would require a substantial sample size to be detected reliably.

Experiment 2

In Experiment 2, we investigated whether the correspondence between racial bias on the IAT and feeling-thermometer measures depends on the stimuli in the explicit measure (verbal category labels vs. images of specific exemplars). Different from the AMP, the standard race IAT includes both images of specific exemplars (i.e., as target stimuli) and verbal category labels (i.e., for the response options). Thus, variance in IAT scores might reflect a mix of responses to images of specific exemplars and responses to verbal category labels. Based on these considerations, we tested whether (1) the association between IAT scores and the explicit category-label measure remains significant when controlling for the explicit exemplar-image measure, and (2) the association between IAT scores and the explicit exemplar-image measure remains significant when controlling for the explicit category-label measure. A significant association between IAT scores and an explicit category-label measure, controlling for an explicit exemplar-image measure, would suggest that responses to verbal category labels play a unique role that is not accounted for by responses to images of specific exemplars. Conversely, a significant association between IAT scores and an explicit exemplar-image measure, controlling for an explicit category-label measure, would suggest that responses to images of specific exemplars play a unique role that is not accounted for by responses to verbal category labels.

Method

Preregistration

The analysis plan was preregistered at https://osf.io/d9q3g/.

Participants

We conducted an a priori power analysis using the Shiny App pwrSEM, which is based on Monte Carlo simulations (Wang & Rhemtulla, 2020). We aimed for at least 80% power to detect a significant association (with α = .05) in two partial regressions of the explicit measures of racial bias on the IAT, as well as model misspecification based on the procedure described by MacCallum et al. (1996). Based on extant reviews (see Gawronski & De Houwer, 2014; Greenwald & Lai, 2020), we assumed the reliability of the explicit category-label measure to be α = .90, of the explicit exemplar-image measure to be α = .80, and of the IAT to be α = .70. Moreover, based on related data reported by Gawronski (2019), the power analysis was based on partial correlations of $r = .12$ between the IAT scores and each of the two explicit measures. We ran 1000 simulations (set seed = 420 and 7) and the app indicated that the minimum target sample size that would provide 80% power to detect all three effects was $N = 545$.[3] We anticipated an exclusion rate of approximately 8% and oversampled to a target sample size of 600, to ensure that we would have at least $N = 545$ for analysis. We preregistered that, if after exclusions and before running any analyses we had a sample size of less than 550, we would compute the exclusion rate (e) and collect $n = (550\text{-current } n)/(1\text{-}e)$ additional participants. We also preregistered that we would exclude participants who (1) showed latencies lower than 300 ms on 10% or more of the trials, or (2) provided a nonsensical response to a Winograd-like attention check designed to filter out bots and inattentive participants.

Participants were Prolific workers over the age of 18, currently living in the USA, fluent in English, and with 80% approval rate who completed the experiment online on Inquisit. A total of 602 rows appear in the raw dataset. Two of these participants were duplicates, and seven had an error in their subject IDs, which made us unable to aggregate their data across tasks. After excluding 8 additional participants who met one or more of our preregistered exclusion criteria ($n = 0$ showed latencies lower than 300 ms on 10% or more of the trials, and $n = 8$ failed the Winograd-like attention check), analyses were conducted on the remaining 585 participants (279 women, 292 men, and 14 people who chose a different option; $M_{\text{age}} = 30.8$, $SD = 11.6$; 65.6% White, 8.5% Black, 12.6% Hispanic or Latino/a/e, 2.7% Black and White Biracial, 5.1% South Asian, 12.5% East Asian, 1.5% American Indian or Alaskan Native, and 3.6% a different identity). A sensitivity analysis conducted on pwrSEM with the same parameter values as above indicated that our final sample provides 83% power to detect a significant path from the explicit category-label measure and 86% power to detect a significant path from explicit exemplar-image measure.

[3] The two partial regressions plus model misspecification were based on the procedure described by MacCallum et al. (1996).

Procedure and Materials

All participants completed a standard race IAT, an explicit category-label measure, and an explicit exemplar-image measure, followed by attention checks and the demographics questions. The order of the three measures of racial bias was counterbalanced across participants. The order of the stimuli was randomized within each of the three bias measures, the only constraint being that faces and adjectives were presented in alternating order in the IAT. We used the labels *Black people* and *White people* as category labels in the IAT and the explicit category-label measure; the adjectives *friendly*, *unfriendly*, *likable*, *dislikable*, *pleasant*, *unpleasant*, *nice*, *nasty*, *good*, and *bad* were used as attribute stimuli in the IAT and for the ratings in the explicit category-label measure. For the exemplar images, we used the same 20 faces of Black and White women and men from Experiment 1.

Explicit Exemplar-Image Measure. The explicit exemplar-image measure was identical to the one in Experiment 1.

Explicit Category-Label Measure. To match the stimuli used in the IAT, the explicit category-label measure consisted of 10 adjective ratings for the categories Black and White (e.g., *White people are unpleasant*). Responses were measured with 7-point scales ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

IAT Measure. The IAT included 80 trials on the compatible and incompatible blocks, respectively, using the faces of the explicit exemplar-image measure as target stimuli and the bipolar adjectives of the explicit category-label measure as attribute stimuli. In the compatible block (same direction as societal bias), participants had to press the *I* key for positive adjectives and White faces, and the *E* key for negative adjectives and Black faces. In the incompatible block, participants had to press the *I* key for positive adjectives and Black faces, and the *E* key for negative adjectives and White faces. The order of the compatible and the incompatible block was counterbalanced across participants. If participants gave an incorrect response, they had to provide the correct response before being able to proceed.

Results

Descriptive statistics of and zero-order correlations between the three measures of racial bias are presented in Table 3. All measures of racial bias showed significant positive associations at the level of zero-order correlations.

For our main analysis, we preregistered that we would use structural-equation modeling (SEM) with latent variables to test (1) the association between the explicit exemplar-image measure and IAT scores while controlling for the explicit category-label explicit measure, and (2) the association between the explicit category-label measure and IAT scores while controlling for the explicit exemplar-image measure.

Following our preregistered analysis plan, we computed 5 parcels for the explicit exemplar-image measure, each reflecing the average of two White faces minus the average of two Black faces. Thus, each parcel is an index of explicit preference for White over Black exemplars. Parcel 1 consisted of White male face 1, White female face 1, Black male face 1 and Black female face 1; parcel 2 consisted of the faces numbered 2, and so on.

For the explicit category-label measure, we computed (1) the difference between ratings of the category White and ratings of the category Black for each of the five positive adjectives (i.e., friendly, likable, pleasant, nice, good), and (2) the difference between ratings of the category Black and ratings of the category White for each of the five negative adjectives (i.e., unfriendly, dislikable, unpleasant, nasty, bad). Next, we computed 5 parcels, each consisting of the average difference score for two adjectives that are direct opposites (e.g., friendly-unfriendly, likeable-dislikeable). Thus, each parcel is an index of explicit preference for the category White over the category Black.

For the IAT, we computed the D-score with built-in error penalty recommended by Greenwald et al. (2003). The two combined blocks were further broken down into 4 sub-blocks with 20 trials each.

The model showed good fit of the data, $\chi^2(74) = 154.25$, $p < .001$, Comparative Fit Index (CFI) = .988, Tucker-Lewis Index (TLI) = .985, Root Mean Square Error of Approximation (RMSEA) = 0.04. The path coefficient for the explicit exemplar-image measure predicting IAT scores was relatively large and statistically significant when controlling for the explicit category-label measure, b(SE) = 0.14(0.04), $p < .001$, $\beta = .31$. In contrast, the path of the explicit category-label measure predicting IAT scores was relatively small and not statistically significant when controlling for the explicit exemplar-image measure, b(SE) = 0.02(0.03), $p = .451$, $\beta = .06$ (see Figure 2).

Discussion

Although racial bias on the IAT, an explicit category-label measure, and an explicit exemplar-image measure were all positively associated at the level of zero-order correlations, an SEM using both explicit exemplar-image and explicit category-label evaluations as predictors of IAT scores revealed a significant path only for the explicit exemplar-image measure after controlling for the explicit category-label measure. The path from the explicit category-label measure was not significant after controlling for the explicit exemplar-image measure. Thus, different from claims that the category labels in the IAT make it uniquely sensitive to category evaluations (e.g., De Houwer, 2001; Fazio & Olson, 2003; Olson & Fazio, 2003), the IAT was highly sensitive to exemplar evaluations, and the association between IAT scores and the explicit category-label measure was fully accounted for by the explicit exemplar-image measure (for related evidence, see Mitchell et al., 2003). Together, these results suggest that, when considering relations between the IAT and explicit measures, the types of stimuli used in the explicit measure (i.e., verbal category labels vs. images of specific exemplars) do matter.

Experiment 3

Experiment 2 provided evidence that stimulus correspondence may indeed matter for associations between explicit and implicit measures. Next, we set out to investigate whether presumed moderators of implicit-explicit relations qualify the correspondence between measures with different stimuli, rather than implicit and explicit measures *per se*. To this end, we chose motivation to control prejudice, which has been found to moderate relations between implicit and explicit measures of bias (e.g., Degner & Wentura, 2008; Dunton & Fazio, 1997; Fazio et al., 1995; Gawronski et al., 2003; Payne et al., 2005). Specifically, participants who report a strong motivation to control prejudice have been found to show a weaker relation between implicit and explicit measures of bias compared to participants who report a weak motivation to control prejudice.

Expanding on the findings of our first two experiments, we investigated whether the obtained moderation pattern is driven by the different types of stimuli rather than the implicit versus explicit nature of the measures. To this end, participants in Experiment 3 completed either an exemplar-image AMP or an exemplar-image feeling thermometer with the same faces. In both conditions, participants additionally completed an explicit category-label measure and the MCPRS (Dunton & Fazio, 1997). We tested whether the relation between the exemplar-image measures and the explicit category-label measure decreases as MCPRS scores increase, and whether this pattern emerges irrespective of whether the exemplar-image measure is implicit or explicit. If the moderation pattern obtained in prior studies is indeed driven by the different types of measures (i.e., implicit vs. explicit) rather than the different types of stimuli (i.e., verbal category labels vs. images of specific exemplars), MCPRS scores should moderate the relation between exemplar-image AMP and the explicit category-label measure, but MCPRS should not moderate the relation between the exemplar-image feeling thermometer and the explicit category-label measure. In contrast, if the moderation pattern obtained in prior studies is driven by the different types of stimuli (i.e., verbal category labels vs. images of specific exemplars) rather than the different types of measures (i.e., implicit vs. explicit), MCPRS scores should moderate not only the relation between exemplar-image AMP and the explicit category-label measure, but also the relation between the exemplar-image feeling thermometer and the explicit category-label measure. Evidence for the latter hypothesis would corroborate the concern that previously obtained findings involving dissociations between explicit and implicit measures may have little to do with the distinction between different types of measures, but may instead be driven by the different stimuli in these measures.

Method

Preregistration

The analysis plan was preregistered at https://osf.io/d9q3g/.

Participants

We preregistered a target sample size of 1080, as indicated by completed surveys on Prolific. We further preregistered that, if after exclusions and before running any analyses we had a sample size of less than 1080, we would compute the exclusion rate (*e*) and collect *n* = (1080-current *n*)/(1-*e*) additional participants. Participants were Prolific workers over the age of 18 who completed the experiment on Inquisit. A total of 1263 rows appear in the raw dataset, out of which 2 were duplicates and 91 had errors in their subject ID that prevented data aggregation.[4] We preregistered that we would exclude participants who (1) provided a nonsensical response to a Winograd-like attention check designed to filter out bots or inattentive participants, and (2) pressed the same key on all trials of the AMP.[5] After excluding the 25 participants who met one or more of these criteria (*n* = 10 failed the Winograd-like attention check, and *n* = 15 pressed the same key), analyses were conducted on the remaining 1145 participants (575 women, 548 men and 22 people who chose a different option; M_{age} = 35.7, *SD* = 12.9; 75.2% White, 9.2% Black, 9.7% Hispanic or Latino/a/e, 2.5% Black and White Biracial, 2.7% South Asian, 6.4% East Asian and 2.7% a different identity). Using

[4] We first collected data from 1080 participants. Out of these, 84 had a subject ID error and 23 were excluded based on our preregistered exclusion criteria, leaving us with a sample of 973. Following our preregistered recruitment plan, we set out to collect data from 118 additional participants. Due to a mistake, 181 were collected instead, out of which 7 had a subject ID error and 2 were excluded based on our preregistered exclusion criteria. Conclusions do not change with or without the extra participants included.

[5] Due to a mistake, we did not preregister that we would exclude participants who knew the meaning of the ideographs. We report results following the preregistered criteria. The conclusions do not change if we exclude participants based on their knowledge of the ideographs.

the zero-order correlations between the variables reported by Payne et al. (2005), a sensitivity power analysis conducted in the InteractionPoweR Shiny App (Baranger et al., 2023) indicated that a sample of 1145 (approximately 573 per condition) provides ~ 94% power to detect an interaction effect of $r = .13$, half the size of the interaction effect reported by Dunton and Fazio (1997).

Procedure and Materials

Participants were randomly assigned to one of two between-subjects conditions in which they completed either an AMP or a feeling thermometer with exemplar images. In addition to completing one of the two exemplar-image measures, all participants completed the MCPRS and an explicit category-label measure. The order of the exemplar-image and category-label measures was counterbalanced, and after participants completed both measures, they completed the MCPRS. The explicit category-label measure, the exemplar-image AMP, and the exemplar-image feeling thermometer were identical to Experiment 1. The MCPRS was directly adapted from Dunton and Fazio (1997). Responses to the 17 items of the MCPRS (e.g., *If I have a prejudiced thought or feeling, I keep it to myself*) were measured with 7-point scales ranging from -3 (*strongly disagree*) to +3 (*strongly agree*). Responses on the MCPRS were averaged to create an index with higher values indicating a stronger motivation to control prejudiced reactions. At the end of the study, participants completed an attention check and demographic questions.

Results

Descriptive statistics of and zero-order correlations between measures are presented in Table 4. Within each condition, MCPRS scores showed significant negative correlations with the measures of racial bias, and each racial-bias measure was posistively correlated with the other racial-bias measures.

Following our preregistered analysis plan, we conducted separate multiple-regression analyses for each condition. In the AMP condition, explicit category-label evaluations were regressed onto standardized MCPRS scores, standardized AMP scores, and the interaction term. The main effect of MCPRS scores was not significant, $\beta = -0.08$, $t(555) = -1.65$, $p = .099$, and the main effect of AMP scores was significant, $\beta = 0.34$, $t(555) = 7.22$, $p < .001$. Critically, the interaction term was significant, $\beta = -0.10$, $t(555) = -2.61$, $p = .009$. Replicating earlier findings, participants with a weak motivation to control prejudiced reactions revealed a stronger positive association between AMP scores and the explicit category-label measure than participants with a strong motivation to control prejudiced reactions (see Figure 3).

In the explicit exemplar-image condition, explicit category-label scores were regressed onto standardized MCPRS scores, standardized explicit exemplar-image scores, and the interaction term. The main effect of MCPRS scores was not significant, $\beta = -0.07$, $t(582) = -1.81$, $p = .071$, and the main effect of explicit exemplar-image scores was significant, $\beta = 0.68$, $t(582) = 17.94$, $p < .001$. Critically, the interaction term was not significant, $\beta = 0.017$, $t(582) = 0.58$, $p = .561$, indicating a positive association between explicit exemplar-image scores and explicit category-label scores regardless of motivation to control prejudiced reactions (see Figure 4).[6]

Discussion

Experiment 1 suggested that the distinction between category-labels and exemplar-images may not be particularly consequential for the AMP, in that the predicted differences were relatively small overall and failed to reach statistical significance in three out of four case. Expanding on these findings, Experiment 3 suggests that the moderating effect of motivation to control prejudiced reactions is limited to the AMP and does not emerge for an explicit exemplar-image measure using the same stimuli.[7] Together, these findings support the conclusion that prior evidence for dissociations between the AMP and explicit category measures may indeed be rooted in the type of measure (i.e., implicit vs. explicit) rather than the type of attitudinal stimuli (i.e., images of specific exemplars vs. verbal category labels).

[6] Although not preregistered, we also tested whether the moderating effect of MCPRS was qualified by the type of exemplar-image measure. To this end, we regressed explicit category-label evaluations onto standardized MCPRS scores, standardized exemplar-image scores, dummy-coded type of exemplar-image measures (implicit = 1 and explicit = 0), and all interactions between the three predictors. Consistent with the conclusion that the moderating effect of MCPRS depended on the type of exemplar measure, the three-way interaction was statistically significant, $\beta = -.12$, $t(1137) = -2.44$, $p = .015$.

[7] Following a reviewer's suggestion, we analyzed the internal consistency of the AMP among participants high vs. low in MCPRS. Interestingly, we found that internal consistency is indeed lower in the high MCPRS group (low: $r = .55$, 95% CI [.46, .62], $p < .001$; high: $r = .42$, 95% CI [.32, .51], $p < .001$; comparison: $z = 1.96$, $p = .0495$), which provides potential alternative account for the differential association between AMP scores among participants with high vs. low scores on the MCPRS. However, we find this same difference for the explicit exemplar-image measure, which also shows lower internal consistency in the high (vs. low) MCPRS group (low: $r = .54$, 95% CI [.46, .62], $p < .001$; high: $r = .34$, 95% CI [.23, .44, $p < .001$; comparison: $z = 3.08$, $p = .002$). If anything, the difference is larger for the explicit exemplar-image measure than the AMP. Thus, if differential reliability drives the MCPRS interaction pattern, we should find the same pattern for the explicit exemplar-image measure, which we do not.

Experiment 4

The goal of Experiment 4 was to investigate the same moderating effect of motivation to control prejudice for the IAT instead of the AMP. To this end, half of the participants completed a standard race IAT; the remaing half completed an explicit exemplar-image measure. As in Experiment 3, participants in both conditions also completed an explicit category-label measure and the MCPRS.

Method

Preregistration

The pre-analysis plan was preregistered at https://osf.io/d9q3g/.

Participants

As in Experiment 3, we preregistered a target sample size of 1080, as indicated by completed surveys on Prolific. We further preregistered that, if after exclusions and before running any analyses we had a sample size of less than 1080, we would compute the exclusion rate (e) and collect n = (1080-current n)/(1-e) additional participants. Participants were Prolific workers over the age of 18 completed the experiment on Inquisit.[8] A total of 1141 rows appear in the raw dataset. Out of these, 30 were duplicates, 2 were incomplete and 20 had a subject ID error so that we could not aggregate their data. We preregistered that we would exclude participants who (1) responded with a latency of less than 300 ms on 10% or more of the trials or (2) provided a nonsensical response to a Winograd-like attention check designed to filter out bots or inattentive participants. After excluding the 10 participants who met one or more of these criteria (n = 4 had 10% of their trials or more with latency of less than 300 ms and n = 6 failed the Winograd-like attention check), analyses were conducted on the remaining 1079 participants (613 women, 431 men and 35 people who chose a different option; M_{age} = 39.9, SD = 14.1; 78.4% White, 6.6% Black, 9.8% Hispanic or Latino/a/e, 3.0% Black and White Biracial, 2.1% South Asian, 6.3% East Asian and 2.7% a different identity). A sensitivity power analysis conducted in the InteractionPoweR Shiny App (Baranger et al., 2023) indicated that a sample of N = 1079 (approximately 540 per condition) provides ~ 93% power to detect an interaction effect of r = .13, half the size of the interaction effect reported by Dunton and Fazio (1997).

Procedure and Materials

As in Experiment 3, participants were randomly assigned to one of two between-subjects conditions in which they completed either the IAT or an explicit exemplar-image measure. All participants additionally completed the explicit category-label measure in counterbalanced order, and finally the MCPRS, which always came last. The explicit category-label measure and explicit exemplar-image measure were identical to Experiment 2. The IAT was identical to Experiment 2, the only difference being that we reduced the number of trials in each of the combined blocks from 80 to 40.[9] At the end of the study, participants completed an attention check and demographic questions.

Results

Descriptive statistics of and zero-order correlations between measures are presented in Table 5. Within each condition, the racial-bias measures were positively correlated with the other racial-bias measures. MCPRS scores showed significant negative correlations with the explicit measures of racial bias, but not with the IAT.

Following our preregistered analysis plan, we conducted separate multiple-regression analyses for each condition.[10] In the IAT condition, explicit category-label scores were regressed onto standardized MCPRS scores, standardized IAT scores, and the interaction term. The main effect of MCPRS was significant, β = -0.31, t(536) = -7.12, p < .001, and the main effect of IAT scores was significant, β = 0.25, t(536) = 5.86, p < .001. Unexpectedly, the interaction term was not significant, β = -0.004, t(540) = -0.09, p = .928, indicating a positive association between IAT scores and the explicit category-label measure regardless of motivation to control prejudiced reactions (see Figure 5).

In the explicit exemplar-image condition, explicit category-label scores were regressed onto standardized MCPRS scores, the standardized explicit exemplar-image scores, and the interaction term. The main effect of MCPRS was not significant, β = -0.02, t(535) = -0.67, p = .502, and the main effect of explicit exemplar-image scores was significant, β = 0.70, t(535) = 18.93, p < .001. Critically, the interaction term was not significant, β = -0.05, t(535) = -1.59, p = .112, indicating a positive association between the explicit exemplar-image measure and the explicit category-

[8] We first collected 1080 participants. Out of these, 20 had a subject ID error and 10 were excluded based on our preregistered exclusion criteria. Following our preregistered recruitment plan, we set out to collect data from 31 additional participants. We collected 30 participants, which due to an error all had duplicate responses; we redid the collection and two had incomplete data.

[9] The reason for this decision was that in Experiment 2 we had to break down the task into parcels to run the SEM. In Experiment 4, there was no need for parcels, so we did not need as many trials.

[10] Two additional preregistered analyses on the factorial structure of the MCPRS are reported in the Supplemental Materials.

label measure regardless of motivation to control prejudiced reactions (see Figure 6).[11]

Discussion

Experiment 4 was set up to test whether both the IAT and the explicit exemplar-image measure would show a moderation by MCPRS when predicting explicit category-label scores. Replicating the findings of Experiment 3, motivation to control prejudiced reactions did not moderate the positive association between racial bias on an explicit exemplar-image measure and racial bias on an explicit category-label measure. Yet, different from the findings with the AMP in Experiment 3, the current study found no significant effect of motivation to control prejudiced reactions on the association between IAT scores and the explicit category-label measure. Instead, the IAT behaved in a manner similar to the explicit exemplar-image measure, which also showed no interaction with motivation prejudiced reactions in the prediction of explicit category-label evaluations. Together with the findings of Experiment 2, the results of Experiment 4 are consistent with the idea that stimulus confounds may indeed matter for findings obtained with the IAT, in that the same findings may be obtained with an explicit exemplar-image measure using the same stimuli. Still, in some cases, the IAT and explicit exemplar-image measures may produce different results, such as the strength of their correlations with other measures (e.g., explicit category-label measures and MCPRS).

General Discussion

Implicit measures often show dissociations from explicit self-report measures, including low correlations between the two kinds of measures (Nosek, 2005), distinct antecedents (Gawronski & Bodenhausen, 2006), and distinct behavioral correlates (Friese et al., 2008). These dissociations are commonly attributed to differences in automaticity features, in that explicit measures capture intentional responses that are relatively easy to control, whereas implicit measures capture unintentional responses that are more difficult to control. However, interpretations of the observed dissociations in terms of automaticity features presuppose that measurement type is not confounded with other important differences. Expanding on concerns that explicit measures of racial bias tend to use verbal category labels whereas many implicit measures of racial bias include images of specific exemplars (Cooley & Payne, 2017; Gawronski, 2019), the current work investigated whether dissociations between implicit and explicit measures of racial bias are at least partly accounted for by the different stimulus materials. To this end, we examined associations between AMP and IAT measures of racial bias with explicit measures using verbal category labels and images of specific exemplars, respectively (Experiments 1 and 2). In addition, we examined for the AMP and the IAT whether motivation to control prejudiced reactions moderates associations between exemplar-image and category-label measures of racial bias regardless of whether the exemplar-image measure is implicit or explicit (Experiments 3 and 4).

Our findings indicate that, although methodologically important, the identified stimulus confound does not seem to matter much for dissociations between the AMP and explicit measures of racial bias. Although correlations between AMP scores and explicit measures of racial bias tended to be larger when the two measures corresponded in terms of their content (i.e., both using images of exemplars or both using verbal category labels) than when they did not correspond (i.e., one using images of exemplars and one using verbal category labels), the obtained differences were very small overall and statistically significant in only one of the four cases (Experiment 1). Moreover, motivation to control prejudiced reactions moderated the association between an exemplar-image AMP and an explicit category-label measure, but motivation to control prejudiced reactions did not show the same moderation effect on the association between an explicit exemplar-image and an explicit category-label measure (Experiment 3). Together, these findings suggest that, although AMP measures of racial bias typically include images of specific exemplars and no verbal category labels, the confound between type of measure and type of stimuli seems to play a relatively minor role for dissociations with explicit measures. Thus, prior evidence for dissociations between the AMP and explicit category-label measures may indeed be rooted in the different types of measures (i.e., implicit vs. explicit) rather than the different types of stimuli (i.e., images of specific exemplars vs. verbal category labels).

Our results were different for the IAT. Counter to claims that the category labels in the IAT make it uniquely sensitive to category evaluations (e.g., De Houwer, 2001; Fazio & Olson, 2003; Olson & Fazio, 2003; but see Mitchell et al., 2003), IAT scores of racial bias showed a significant positive association with an explicit exemplar-image measure after controlling for an explicit category-label measure, but IAT scores of

[11] Although not preregistered, we also regressed explicit category evaluations onto standardized MCPRS scores, standardized exemplar evaluations, dummy-coded type of exemplar measures (i.e., implicit = 1 vs. explicit = 0), and all interactions between the three predictors. The three-way interaction was not statistically significant, $\beta = .04$, $t(1071) = 0.81$, $p = .419$, suggesting that the moderating effect of MCPRS did not depend on the type of exemplar measure.

racial bias were unrelated to an explicit category-label measure after controlling for an explicit exemplar-image measure (Experiment 2). Even though the experiment does not speak to the reason behind the unique shared variance between the exemplar-image measure and the IAT, the result suggests that the types of stimuli used matter. Moreover, different from the results obtained with the AMP, motivation to control prejudiced reactions did not moderate the association between IAT scores and an explicit category-label measure (Experiment 4). Instead, racial bias on the IAT showed a small but statistically significant positive association with an explicit category-label measure regardless of motivation to control prejudiced reactions. Thus, across studies, racial bias on the IAT showed strong correspondence (Experiment 2) and functional resemblance (Experiment 4) with racial bias on an explicit exemplar-image measure. While inferences from these findings are partly based on an unexpected null effect (i.e., non-significant moderation effect of motivation to control prejudice in Experiment 4), the overall pattern of results is consistent with the idea that the identified stimulus confound does matter for the IAT. Thus, IAT measures of racial bias may show dissociations with explicit category-label measures because of the exemplar images in the IAT rather than the implicit versus explicit nature of the measurement instruments.

Relation to Prior Findings

An important question is how the current findings can be reconciled with earlier findings showing that associations between implicit and explicit measures of bias tend to be stronger for participants with a weak motivation to control prejudiced reactions compared to participants with a strong motivation to control prejudiced reactions (e.g., Akrami & Ekehammar, 2005; Degner & Wentura, 2008; Dunton & Fazio, 1997; Fazio et al., 1995; Gawronski et al., 2003; Payne et al., 2005). In the current work, we replicated this widely cited pattern for the AMP but not for the IAT. Regarding the null effect obtained for the IAT, it is worth noting that prior studies on the presumed effect of motivation to control prejudiced reactions used sample sizes that, by today's standards, may be deemed insufficient for the detection of the hypothesized two-way interaction (da Silva Frost & Ledgerwood, 2020), with samples sizes between *N* = 42 (Akrami & Ekehammar, 2005) and *N* = 111 (Fazio et al., 1995). We are aware of five published studies that found a significant interaction between IAT-measured evaluative bias and MCPRS scores in the prediction of explicit category-label evaluations, and all these studies seem underpowered for the detection of the predicted two-way interaction with sample sizes of *N* = 42 (Akrami & Ekehammar, 2005), *N* = 69 (Gawronski et al., 2003), *N* = 87 (Hofmann, Gschwendner, et al., 2005, Study 2), *N* = 93 (Hofmann, Gschwendner, et al., 2005, Study 1), and *N* = 103 (Ziegert & Hanges, 2005), respectively. Although we cannot rule out that the non-significant interaction between MCPRS scores and IAT-measured racial bias in Experiment 4 reflects a false negative, the large sample size in that study (*N* = 1079) renders such an interpretation unlikely. Instead, it seems more likely that prior findings suggesting an interaction between IAT-measured evaluative bias and MCPRS scores in the prediction of explicit category-label evaluations are false positives (see Button et al., 2013), or that methodological differences between the experiments led to discrepant results. Future research using large samples and diverse methods may help to resolve this question.

The current findings expand on prior work addressing issues closely related to the identified stimulus confound. For example, Cooley and Payne (2017) noted that many implicit measures of racial bias capture responses to individual exemplars, and that a person's aggregated responses to individual exemplars may not correspond to the person's response when encountering the same exemplars in a group. To address this issue, Cooley and Payne (2017) presented a modified variant of the AMP that uses images of groups as prime stimuli (e.g., an image showing three Black men) instead of individual exemplars. Across a series of studies, the authors found systematic differences between racial bias on the traditional exemplar-image AMP and racial bias on the newly developed group-image AMP (see also Cooley & Payne, 2019). The current findings expand on this work by showing that confounds between type of measure (i.e., implicit vs. explicit) and type of stimuli (i.e., images of exemplars vs. verbal category labels) play a relatively minor role for dissociations between traditional exemplar-image AMPs and explicit category-label measures. However, this conclusion does not question Cooley and Payne's (2017) finding that aggregated responses to individual exemplars can differ from responses to a group comprising the same exemplars.

The current research also expands on prior work by Payne et al. (2008), which in our view provides the most comprehensive analysis of confounds between type of measure and other characteristics. In addition to the stimulus confound addressed in the current studies, Payne et al. aimed to control for various other structural confounds, such as different response formats (e.g., rating scale vs. binary categorization) and response type (e.g., evaluative ratings vs. response times). To this end, the authors utilized two matched variants of the AMP that differed only in terms of whether the measure was implicit or explicit. In the traditional implicit variant, participants were instructed to evaluate the Chinese ideographs and to ignore the primes. In the newly developed explicit variant, participants were instructed to evaluate the primes and to ignore the Chinese ideographs. A major advantage of this approach is that it controls for various other confounds beyond the stimulus confound addressed in the current work (see Corneille & Gawronski, 2024). Yet, a notable disadvantage is that, by controlling all of these confounds, it is not possible to gauge the relative impact of each individual confound, including the confound between type of measure and type of stimuli. Moreover, because Payne et al.'s structural-fit approach is applicable only to the AMP, it is not suitable to gauge the role of stimulus confounds in other implicit measures such as the IAT. The current findings suggest that, if anything, confounds between type of measure (i.e., implicit vs. explicit) and type of stimuli (i.e., images of specific exemplars vs. verbal category labels) play a more significant role in the IAT compared to the AMP. Nevertheless, we deem the two AMP variants presented by Payne et al. (2008) an ideal approach to resolving multiple confounds, including the stimulus confound addressed in the current work.

Recommendations

The framework and findings described in this article have important implications for racial-bias research comparing responses on implicit and explicit measures. Dissociations between the two kinds of measures are typically interpreted in terms of automaticity features, in that implicit measures are assumed to capture unintentional responses that are relatively difficult to control whereas explicit measures capture intentional responses that are relatively easy to control. However, such interpretations require that implicit and explicit measures do not differ in terms of other features, such as the stimuli in the two kinds of measures. While this issue has been long acknowledged in research using implicit and explicit measures to study personality self-concepts (e.g., Asendorpf et al., 2002; Back et al., 2009; Peters & Gawronski, 2011), it has been largely ignored in research on racial bias, where implicit measures commonly include images of specific exemplars while explicit measures predominantly rely on verbal category labels without presenting images of specific exemplars (for discussions, see Cooley & Payne, 2017; Gawronski, 2019; Payne et al., 2008). The current findings suggest that, while the confound between type of measure and type of stimuli seems to play a relatively minor role for the AMP, it seems potentially important for the IAT. Hence, prior findings showing low correlations between IAT measures of social bias and explicit self-report measures (for a meta-analysis, see Hofmann, Gawronski, et al., 2005), distinct antecedents (for a meta-analysis, see Forscher et al., 2019), and distinct behavioral correlates (for a meta-analysis, Kurdi et al., 2019) may have nothing to do with commonly invoked difference between types of measures (i.e., implicit vs. explicit). Instead, the same dissociations may emerge for explicit measures using exemplar-images, which would suggest fundamentally different conclusions for both theory and practice. Based on these considerations, we recommend that future research using implicit and explicit measures of social bias should always include an explicit exemplar-image measure in addition to the commonly used explicit category-label measures. Such designs permit stronger conclusions about whether dissociations are driven by the different types of measures or the different types of stimuli.

In our view, the best approach to addressing all confounds in comparisons between implicit and explicit measures is Payne et al.'s (2008) structural-fit approach, using two matched variants of the AMP that differ only in terms of whether participants are instructed to evaluate the prime stimuli (e.g., Black and White faces) or the target stimuli (e.g., Chinese ideographs). Because the two AMP variants differ only in terms of whether the prime stimuli influence responses intentionally or unintentionally, dissociations between two AMP variants provide more solid evidence for inferences about unintentionality compared to dissociations between implicit and explicit measures that also differ in terms of other features (e.g., types of stimuli). While some previously obtained dissociations between implicit and explicit measures have been replicated with the two variants of the AMP (e.g., Payne et al., 2008), other dissociations disappeared in that the two AMP variants showed identical effects (e.g., Béna et al., 2022). Based on these findings and the results of the current work, we recommend using Payne et al.'s (2008) structural-fit approach instead of comparing responses on implicit and explicit measures that differ in multiple ways beyond their implicit and explicit nature (see also Corneille & Gawronski, 2024).

Constraints on Generality

Despite several important strengths (e.g., carefully controlled experimental setting, preregistration, large sample sizes), the current work also has some notable limitations. First, the current work focused exclusively on the AMP and the IAT, which are only two instruments among the large set of the currently available implicit measures (for reviews, see Gawronski & De Houwer, 2014; Greenwald & Lai, 2020). In addition to their greater prominence, our focus on the AMP and the IAT was based on the facts that (1) high internal consistency is methodologically imperative for the correlational designs of the current studies (see Koppehele-Gossel et al., 2020) and (2) the AMP and the IAT are the only two implicit measures that meet this criterion (see Gawronski & De Houwer, 2014; Greenwald & Lai, 2020). Nevertheless, future research investigating the confound between type of measure and type of stimuli for other implicit measures would be extremely valuable.

Second, the current work focused exclusively on measures of racial bias involving evaluative responses to Black and White targets. Yet, social biases exist for a broad range of groups, and not all of these biases involve evaluative responses (Amodio & Devine, 2006). Because the confound between type of measure and type of stimuli seems relevant for all of these cases, future studies on social biases against other groups involving non-evaluative dimensions (e.g., semantic gender stereotypes) would be helpful to address the generality of our findings.

Finally, all of the reported studies have been conducted with participants from the United States, which raises the question of whether the obtained results would replicate in samples from other countries. For example, some researchers have argued that racial categories tend to be more salient in the United States compared to many European countries (e.g., Degner & Wentura, 2010), which may affect the interplay of racial bias at the level of specific exemplar images and abstract category labels in either explicit or implicit measures (or both). Thus, future research investigating the reproducibility of our findings with samples of non-American participants would be helpful to gauge the generalizability of our conclusions.

Conclusion

The current research addressed a common confound in research on racial bias: the confound between type of measure (i.e., implicit vs. explicit) and type of stimuli (i.e., images of specific exemplars vs. verbal category labels). Our findings suggest that, while the identified confound does not seem to matter much for the AMP, it does seem to matter for the IAT in that the IAT showed strong correspondence and functional resemblance to an explicit measure using images of specific exemplars. These findings raise important questions about whether previously obtained dissociation between self-report and IAT measures of racial bias are driven by the explicit versus implicit nature of the instruments or the different stimuli in the two kinds of instruments. Based on these conclusions, we recommend that future research should control for stimulus confounds when comparing racial bias on implicit and explicit measures.

Author Note

This research was supported by National Science Foundation Grant BCS-1941440. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Aline da Silva Frost contributed with Methodology, Formal Analysis, Investigation, Data Curation, Project Administration, Writing – Original Draft, and Writing – Review and Editing. Alison Ledgerwood contributed with Conceptualization, Methodology, Writing - Review and Editing, Supervision, and Funding acquisition. Paul Eastwick contributed with Conceptualization, Methodology, Writing - Review and Editing, Supervision, and Funding acquisition. Bertram Gawronski contributed with Conceptualization, Methodology, Writing - Original Draft, and Funding acquisition.

References

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84*, 888-918.

Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracın, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 173–221). Mahwah, NJ: Erlbaum.

Akrami, N., & Ekehammar, B. (2005). The association between implicit and explicit prejudice: The moderating role of motivation to control prejudiced reactions. *Scandinavian Journal of Psychology*, *46*, 361-366.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91*, 652–661.

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology, 83*, 380-393.

Axt, J. R. (2018). The best way to measure explicit racial attitudes is to ask about them. *Social Psychological and Personality Science*, *9*, 896-906.

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology, 97*, 533-548.

Baranger, D. A., Finsaas, M., Goldstein, B., Vize, C., Lynam, D., & Olino, T. M. (2023). Tutorial: Power analyses for interaction effects in cross-sectional regressions. *Advances in Methods and Practices in Psychological Science, 6,* 1-13.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 1-40). Hillsdale, NJ: Erlbaum.

Béna, J., Melnikoff, D. E., Mierop, A., & Corneille, O. (2022). Revisiting dissociation hypotheses with a structural fit approach: The case of the prepared reflex framework. *Journal of Experimental Social Psychology*, *100*, Article 104297.

Button, K., Ioannidis, J., Mokrysz, C., Nosek, B., Flint, J., Robinson, E., & Munafo, M. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376

Cameron, C. D., Brown-Iannuzzi, J., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behaviors and explicit attitudes. *Personality and Social Psychology Review, 16,* 330-350.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cooley, E., & Payne, B. K. (2017). Using groups to measure intergroup prejudice. *Personality and Social Psychology Bulletin, 43,* 46-59.

Cooley, E., & Payne, B. K. (2019). A group is more than the average of its parts: Why existing stereotypes are applied more to the same individuals when viewed in groups than when viewed alone. *Group Processes & Intergroup Relations, 22*, 673-687.

Corneille, O., & Gawronski, B. (2024). Self-reports are better measurement instruments than implicit measures. *Nature Reviews Psychology, 3,* 835-846.

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review*, *24*, 212-232.

da Silva Frost, A., & Ledgerwood, A. (2020). Calibrate your confidence in research findings: A tutorial on improving research methods and practices. *Journal of Pacific Rim Psychology, 14,* Article e14.

da Silva Frost, A., Wang, Y. A., Eastwick, P. W., & Ledgerwood, A. (2024). Summarized attribute preferences have unique antecedents and consequences. *Journal of Experimental Psychology: General, 153,* 913-938.

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*, *37*, 443-451.

De Houwer, J., & Boddez, Y. (2022). Bias in implicit measures as instances of biased behavior under suboptimal conditions in the laboratory. *Psychological Inquiry*, *33*, 173-176.

Degner, J., & Wentura, D. (2008). The extrinsic affective Simon task as an instrument for indirect assessment of prejudice. *European Journal of Social Psychology, 38,* 1033-1043.

Degner, J., & Wentura, D. (2010). Automatic prejudice in childhood and early adolescence. *Journal of Personality and Social Psychology, 98*, 356-374.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23,* 316-326.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition & Emotion*, *15*, 115-141.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*, 1013–1027.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*, 297-327.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology, 117*, 522-559.

Friese, M., Hofmann, W., & Schmitt, M. (2008). When and why do implicit measures predict behaviour? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology*, *19*, 285-338.

Fujita, K., Eyal, T., Chaiken, S., Trope, Y., & Liberman, N. (2008). Influencing attitudes toward near and distant objects. *Journal of Experimental Social Psychology*, *44*, 562-572.

Fujita, K., & Han, H. A. (2009). Moving beyond deliberative control of impulses: The effect of construal levels on evaluative associations in self-control conflicts. *Psychological Science, 20,* 799-804.

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14,* 574-595.

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.

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* (2nd edition, pp. 283-310). New York: Cambridge University Press.

Gawronski, B., De Houwer, J., & Sherman, J. W. (2020). Twenty-five years of research using implicit measures. *Social Cognition, 38,* s1-s25.

Gawronski, B., Deutsch, R., LeBel, E. P., & Peters, K. R. (2008). Response interference as a mechanism underlying implicit measures: Some traps and gaps in the assessment of mental associations with experimental paradigms. *European Journal of Psychological Assessment*, *24,* 218-225.

Gawronski, B., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology*, *33*, 573-589.

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022a). Implicit bias ≠ bias on implicit measures. *Psychological Inquiry*, *33*, 139-155.

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022b). Reflections on the difference between implicit bias and bias on implicit measures. *Psychological Inquiry*, *33*, 219-231.

Gawronski, B., & Ye, Y. (2014). What drives priming effects in the affect misattribution procedure? *Personality and Social Psychology Bulletin*, *40*, 3-15.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4-27.

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology*, *71*, 419-445.

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.

Greenwald, A. G., & Nosek, B. A. (2009) Attitudinal dissociation: What does it mean? In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 65-82). New York: Psychology Press.

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

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*, 1369-1392.

Hansen, J., & Trope, Y. (2013). When time flies: How abstract and concrete mental construal affect the perception of time. *Journal of Experimental Psychology: General, 142*, 336-347.

Henderson, M. D. (2013). When seeing the forest reduces the need for trees: The role of construal level in attraction to choice. *Journal of Experimental Social Psychology*, *49*, 676-683.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*, 1369-1385.

Hofmann, W., Gschwendner, T., & Schmitt, M. (2005). On implicit–explicit consistency: The moderating role of individual differences in awareness and adjustment. *European Journal of Personality*, *19*, 25-49.

Irving, L. H., & Smith, C. T. (2020). Measure what you are trying to predict: Applying the correspondence principle to the Implicit Association Test. *Journal of Experimental Social Psychology*, *86*, Article 103898.

Koppehele-Gossel, J., Hoffmann, L., Banse, R., & Gawronski, B. (2020). Evaluative priming as an implicit measure of evaluation: An examination of outlier-treatments for evaluative priming scores. *Journal of Experimental Social Psychology, 87,* Article 103905.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist, 74*, 569-586.

Ledgerwood, A., Eastwick, P. W., & Smith, L. K. (2018). Toward an integrative framework for studying human evaluation: Attitudes towards objects and attributes. *Personality and Social Psychology Review, 22*, 378-398.

Ledgerwood, A., Trope, Y., & Chaiken, S. (2010). Flexibility now, consistency later: Psychological distance and construal shape evaluative responding. *Journal of Personality and Social Psychology, 99*, 32-51.

Ledgerwood, A., Wakslak, C. J., Sánchez, A. M., & Rees, H. R. (2019). A brief, distance-based intervention can increase intentions to follow evidence-based guidelines in cancer screening. *Social Psychological and Personality Science*, *10*, 653-661.

Ledgerwood, A., & Wang, Y. A. (2018). Achieving local and global shared realities: Distance guides alignment to specific or general social influences. *Current Opinion in Psychology*, *23*, 62-65.

Levesque, H. J., Davis, E., & Morgenstern, L. (2012, June). The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning* (pp. 552-561).

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

McCrea, S. M., Wieber, F., & Myers, A. (2012). Construal level mind-sets moderate self- and social stereotyping. *Journal of Personality and Social Psychology, 102*, 51–68.

Melnikoff, D. E., & Kurdi, B. (2022). What implicit measures of bias can do. *Psychological Inquiry*, *33*, 185-192.

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, *36*, 630-633.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, *132*, 455-469.

Moors, A., & De Houwer, J. (2006). Automaticity: A conceptual and theoretical analysis. *Psychological Bulletin, 132*, 297-326.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*, 565-584.

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences*, *15*, 152-159.

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test. *Experimental Psychology*, *54*, 14-29.

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science, 14*, 636-639.

Olson, M. A., & Gill, L. J. (2022). Commentary on Gawronski, Ledgerwood, and Eastwick, Implicit Bias ≠ Bias on Implicit Measures. *Psychological Inquiry*, *33*, 199-202.

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology, 94*, 16-31.

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.

Payne, B. K., & Lundberg, K. (2014). The affect misattribution procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass, 8*, 672-686.

Peters, K. R., & Gawronski, B. (2011). Mutual influences between the implicit and explicit self-concepts: The role of memory activation and motivated reasoning. *Journal of Experimental Social Psychology*, *47*, 436-442.

Ratliff, K. A., & Smith, C. T. (2022). Implicit bias as automatic behavior. *Psychological Inquiry*, *33*, 213-218.

Rim, S., Amit, E., Fujita, K., Trope, Y., Halbeisen, G., & Algom, D. (2015). How words transcend and pictures immerse: On the association between medium and level of construal. *Social Psychological and Personality Science*, *6*, 123-130.

Soderberg, C. K., Callahan, S. P., Kochersberger, A. O., Amit, E., & Ledgerwood, A. (2015). The effects of psychological distance on abstraction: Two meta-analyses. *Psychological Bulletin, 141,* 525-548.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*, 440-463.

Wakslak, C., & Trope, Y. (2009). The effect of construal level on subjective probability estimates. *Psychological Science, 20*, 52-58.

Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, *4*, 1-17.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology, 72*, 262–274.

Ziegert, J. C., & Hanges, P. J. (2005). Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology, 90*, 553–562.

Table 1
Descriptive Statistics of Racial-Bias Scores as a Function of Measurement Type (Explicit vs. Implicit) and Stimulus Type (Exemplar-Image vs. Category-Label), Experiment 1

Measure	*n*	*M*	*SD*	*α*
Explicit Category-Label	611	0.21	1.22	.69
Explicit Exemplar-Image	510	-0.01	1.01	.91
Implicit Category-Label	564	0.02	0.29	.72
Implicit Exemplar-Image	557	0.02	0.27	.65

Table 2
Correlations Between Measures of Racial Bias as a Function of Measurement Type (Explicit vs. Implicit) and Measurement Content (Exemplar-Image vs. Category-Label), Experiment 1

Condition	*n*	*r*	95% CI	*p*
A. Explicit Exemplar - Implicit Exemplar	264	.50	[.40, .58]	< .001
B. Explicit Exemplar - Implicit Category	246	.41	[.30, .51]	< .001
C. Explicit Category - Implicit Category	318	.56	[.48, .63]	< .001
D. Explicit Category - Implicit Exemplar	293	.45	[.35, .54]	< .001

Table 3
Descriptive Statistics of and Zero-Order Correlations Between Measures, Experiment 2

Measure	*n*	*M*	*SD*	α	1	2	3
1. Explicit Category-Label	585	-0.50	1.16	.94	-		
2. Explicit Exemplar-Image	585	-0.35	1.02	.87	.71***	-	
3. IAT	585	0.33	0.42	.95	.29***	.33***	-

Note: * *p* < .05, ** *p* < .01, *** *p* < .001

Table 4
Descriptive Statistics of and Zero-Order Correlations Between Measures as a Function of Measurement Condition (Explicit Exemplar-Image vs. AMP), Experiment 3

Measure	*M*	*SD*	*α*	1	2	3
Explicit Exemplar-Image (*n* = 586)						
Explicit Category-Label	0.03	1.10	.68	-		
MCPRS	0.50	0.80	.87	-.15***	-	
Explicit Exemplar-Image	-0.22	0.88	.92	.62***	-.15***	-
AMP (*n* = 559)						
Explicit Category-Label	0.06	1.13	.64	-		
MCPRS	0.56	0.83	.88	-.11*	-	
AMP	0.16	0.26	.68	.33***	-.11**	-

Note: * *p* < .05, ** *p* < .01, *** *p* < .001

Table 5

Descriptive Statistics of and Zero-Order Correlations Between Measures as a Function of Measurement Condition (Explicit Exemplar-Image vs. IAT), Experiment 4

Measure	*M*	*SD*	*α*	1	2	3
Explicit Exemplar-Image (*n* = 539)						
1. Explicit Category-Label	-0.47	1.10	.93	-		
2. MCPRS	0.54	0.90	.90	-.14***	-	
3. Explicit Exemplar-Image	-0.24	0.84	.90	.65***	-.19***	-
IAT (*n* = 540)						
1. Explicit Category-Label	-0.42	1.07	.95	-		
2. MCPRS	0.55	0.83	.87	-.29***	-	
3. IAT	0.38	0.39	.66	.24***	-.005	-

Note: * *p* < .05, ** *p* < .01, *** *p* < .001

Figure 1

Illustration of Potential Associations Between Measures of Racial Bias as a Function of Measurement Type (Implicit vs. Explicit) and Measurement Content (Exemplar Images vs. Category Labels)

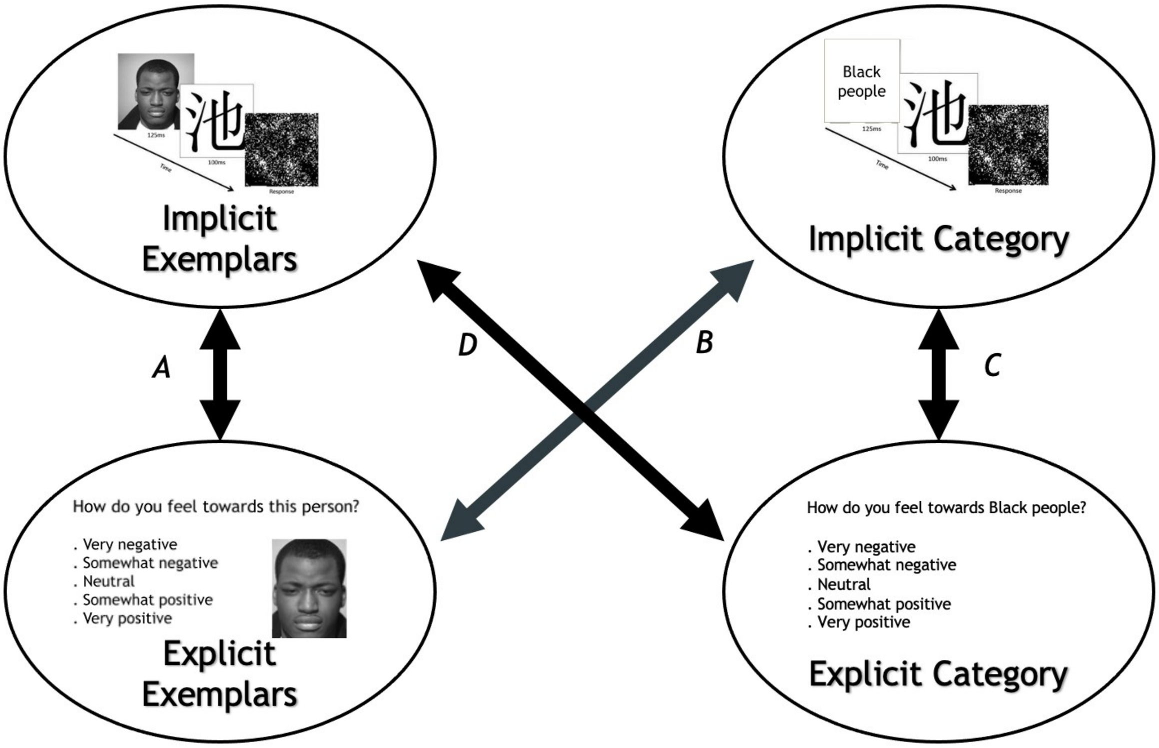


Figure 2

Results of Structural Equation Modeling Predicting Racial Bias on the IAT via Explicit Exemplar-Image Evaluations and Explicit Category-Label Evaluations

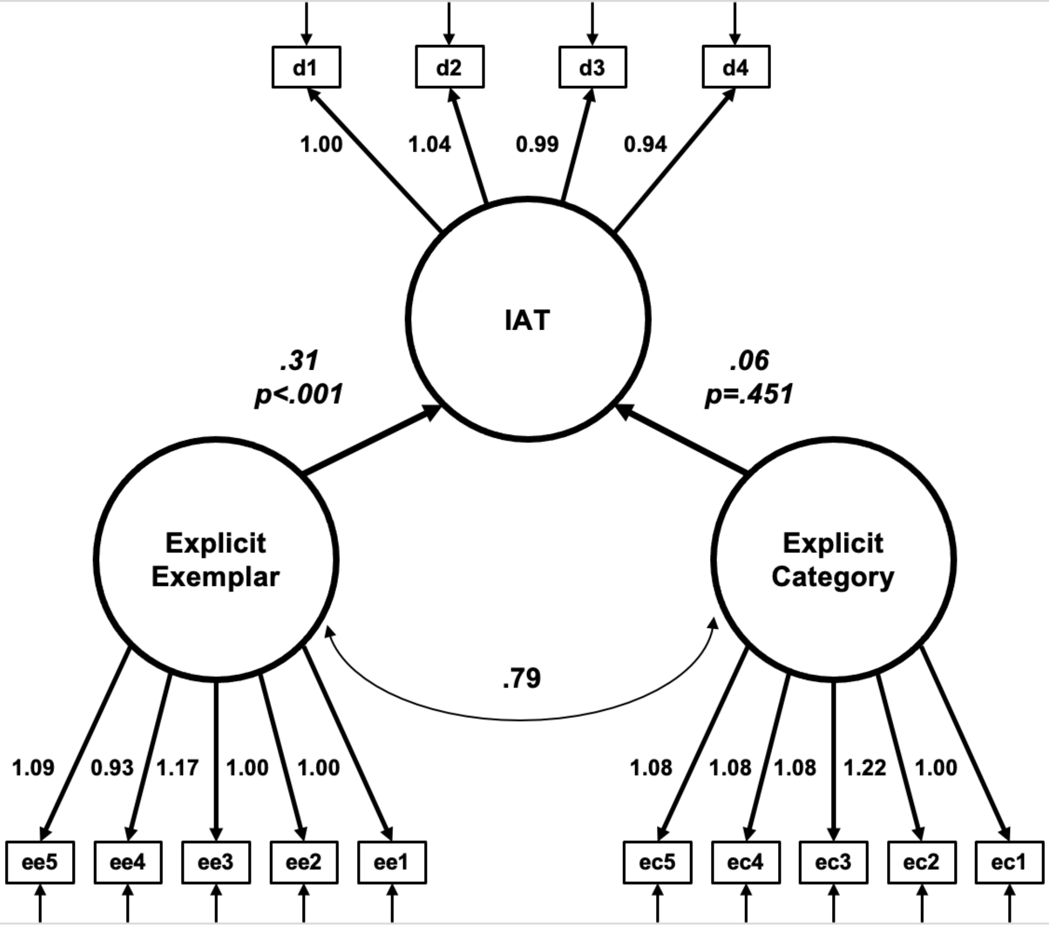


Figure 3

Explicit Category-Label Evaluations as a Function of Racial Bias on the AMP and Motivation to Control Prejudiced Reactions (MCPRS), Experiment 3

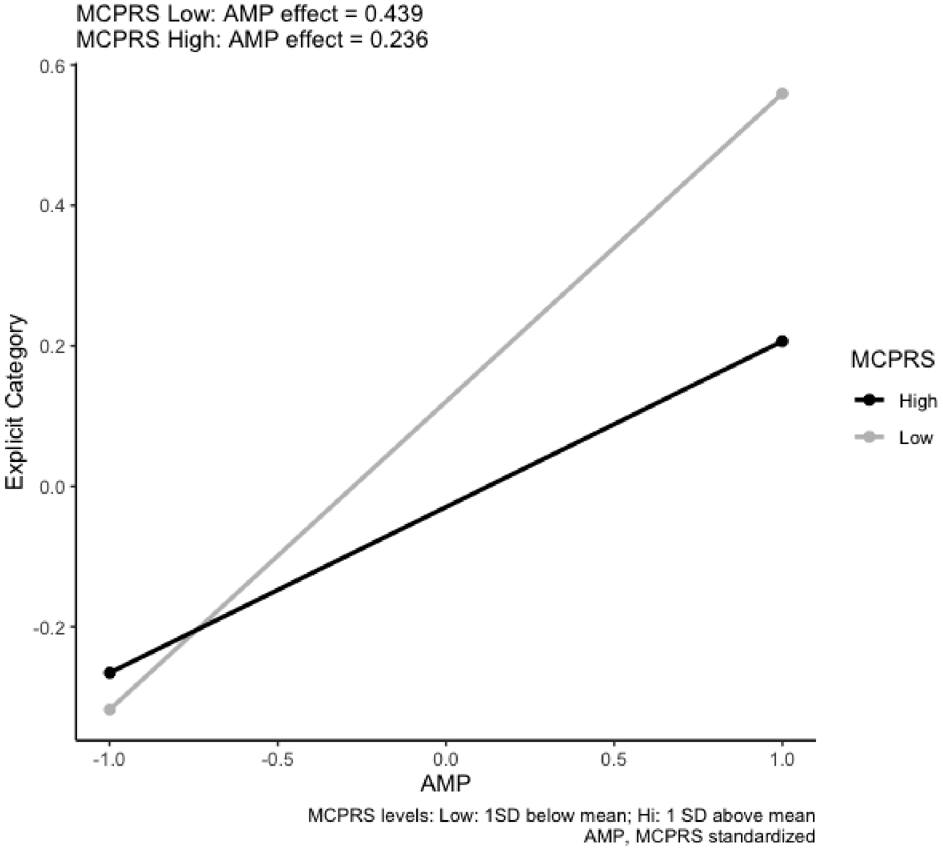


Figure 4

Explicit Category-Label Evaluations as a Function of Explicit Exemplar-Image Evaluations and Motivation to Control Prejudiced Reactions (MCPRS), Experiment 3

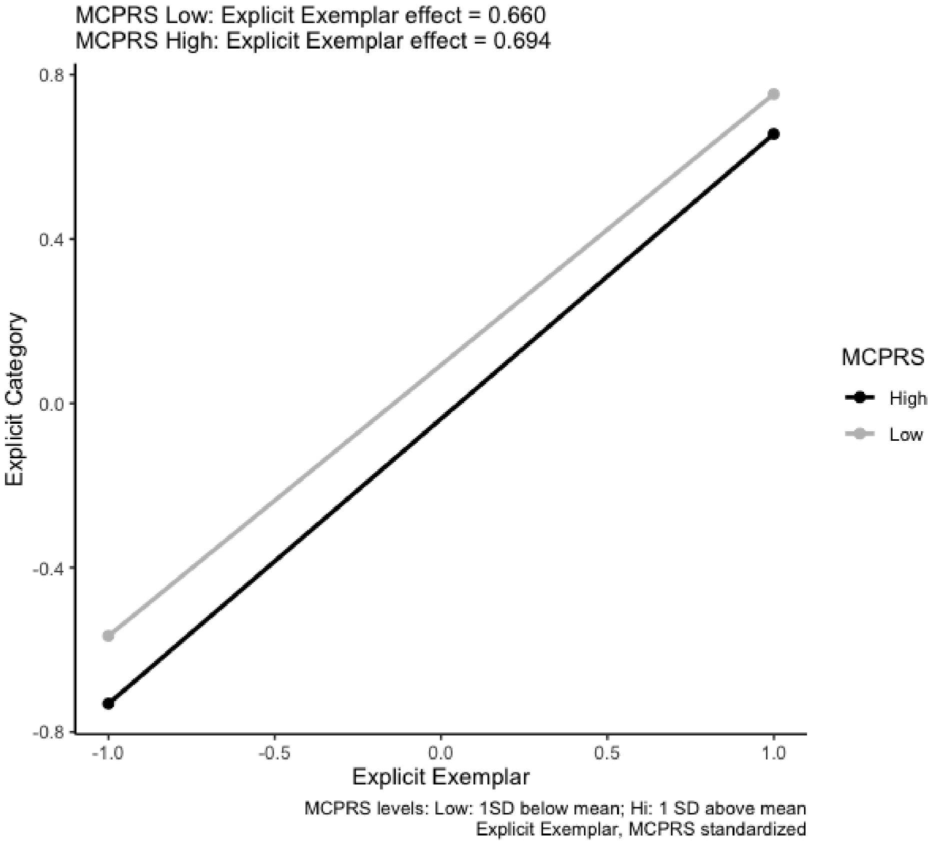


Figure 5

Explicit Category-Label Evaluations as a Function of Racial Bias on the IAT and Motivation to Control Prejudiced Reactions (MCPRS), Experiment 4

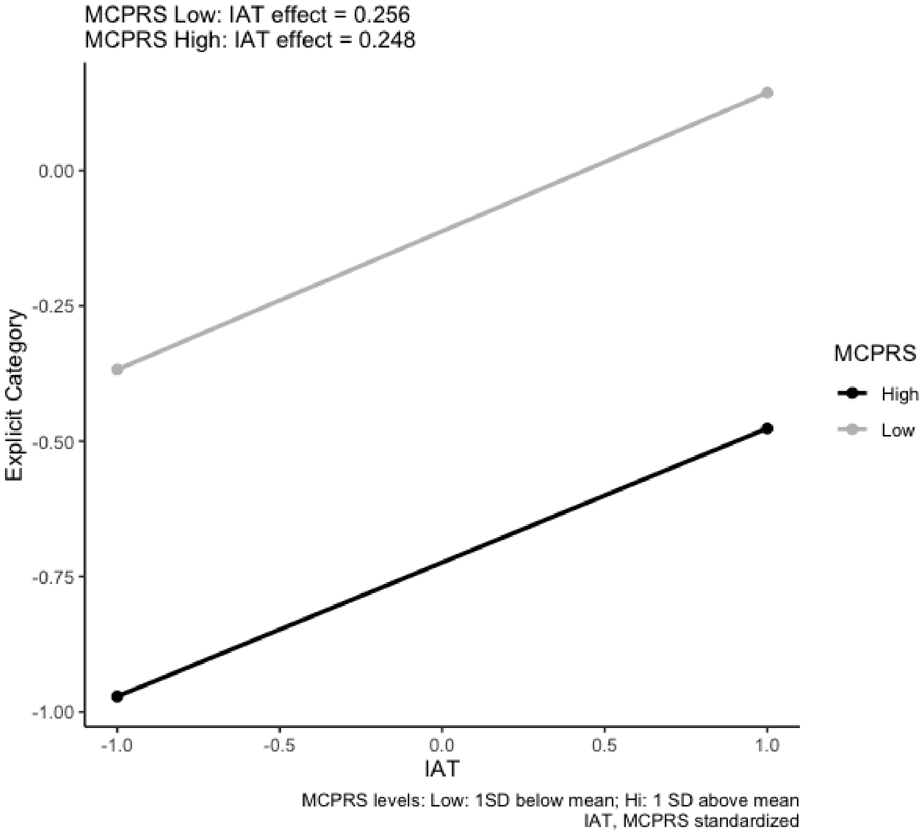
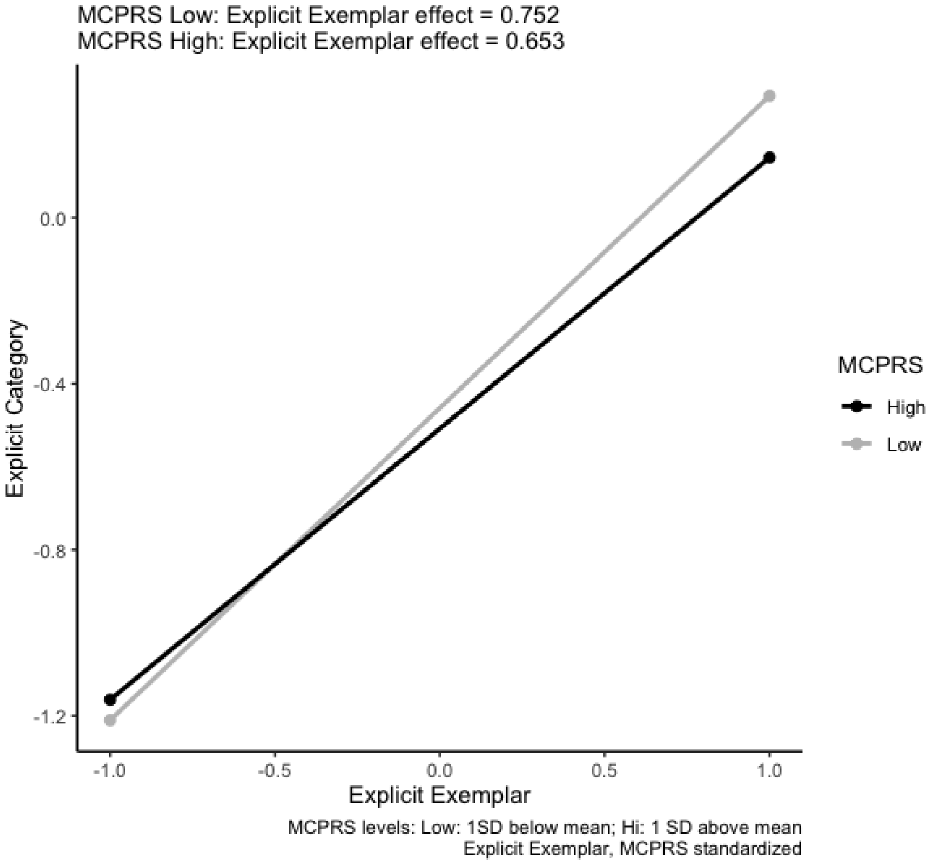


Figure 6

Explicit Category-Label Evaluations as a Function of Explicit Exemplar-Image Evaluations and Motivation to Control Prejudiced Reactions (MCPRS), Experiment 4


MCPRS Low: Explicit Exemplar effect = 0.752
MCPRS High: Explicit Exemplar effect = 0.653
Explicit Category
Explicit Exemplar
MCPRS
High
Low
MCPRS levels: Low: 1SD below mean; Hi: 1 SD above mean
Explicit Exemplar, MCPRS standardized