

A Formal Mathematical Examination of Egalitarianism in the Implicit Association Test (IAT)

Liz Wilson

University of California, Riverside; Harvard University

Ruben Laukenmann

University of Mannheim; University of St. Gallen

Jimmy Calanchini

University of California, Riverside

Past research using the quad model indicates that multiple cognitive processes underlie responses on the Implicit Association Test (IAT). The quad model assumes that one target group is evaluated positively (e.g., White-good) and the other target group is evaluated negatively (e.g., Black-bad). Though theoretically defensible, this assumption precludes the possibility that both target groups are evaluated positively, as some theories of intergroup bias posit. We sought to validate an egalitarian extension of the quad model (E-quad model) that estimates positive evaluations of both target groups. In a series of six experiments, we found that the E-quad model fits IAT data, but we did not find consistent evidence of the model's construct validity. Moreover, model selection indices suggest that the original quad model fits these data better than does the E-quad model. Taken together, these experiments ultimately support the quad model as better for the IAT.

Keywords: formal mathematical models, implicit measure, quad model

Though the United States has made many strides forward in advancing issues of social justice and equity, minoritized racial groups still experience social disadvantages and persistent disparities in areas ranging from health, workplace settings, policing, and education, among many other domains (Dovidio et al., 2008; K.P. Jones et al., 2017; S.C.T. Jones et al., 2020; Levy et al., 2016; Pierson et al., 2020; Riddle & Sinclair, 2019). These disparities have inspired lines of research that interrogate intergroup bias, especially in its contemporary forms, which often occur in relatively subtle and pernicious ways (Oswald et al., 2013; Sue et al., 2007). In light

of concerns that people are sometimes unable or unwilling to accurately report their evaluations of social groups, researchers developed bias measures aimed to constrain the influence of introspection failures and social desirability biases. Accordingly, the use of measures designed to capture these implicit¹ biases has proliferated over the past few decades, and some evidence suggests they reflect mental processes that operate unintentionally (Fazio et al., 1995; Mann et al., 2019; Melnikoff & Kurdi, 2022).

The most widely used measure of implicit bias is the Implicit Association Test (IAT: Greenwald et al., 1998), and a variety of social cognitive theories assume that such measures primarily assess mental association stored in memory (Fazio, 1990; Petty et al., 2007; Strack & Deutsch, 2004). The IAT is a speeded dual categorization task in which people respond to stimuli reflecting two target categories (e.g., White, Black) and two attributes (e.g., good, bad). Responses on the race IAT have traditionally been quantified according to the *D*-scoring algorithm (Greenwald et al., 2003), a summary statistic that reflects the standardized difference between participants' mean response latencies to normatively compatible trials² (i.e., when White stimuli share a response key with good attributes) versus incompatible trials (i.e., when Black stimuli share a response key with good attributes). On an IAT configured to assess Black/White implicit racial bias, positive *D*-scores are generally interpreted to reflect relatively more positive evaluations of White people, and negative *D*-scores reflect relatively more positive evaluations of Black people. However, operationalizing IAT responses as a relative summary statistic is theoretically imprecise and precludes the possibility that multiple processes contribute to IAT responses.

Multinomial processing tree (MPT: Riefer & Batchelder, 1988) models serve as an alternative IAT scoring approach to the *D*-score. Rather than summarizing responses into a single index, MPT models independently quantify the joint contributions of multiple cognitive processes to responses, such as evaluations and inhibition. This theoretical contribution of MPT modeling is crucial because previous research has demonstrated that *D*-scoring versus MPT modeling of IAT responses can imply different interpretations of results. For example, older people who complete a race IAT often produce *D*-scores that would seem to reflect greater pro-White/anti-Black bias compared to the *D*-scores of younger people (Nosek et al., 2002). One interpretation of these findings is that older people are more racially biased because they grew up in times when prejudice was more prevalent and acceptable. However, one MPT model, the quad model (Conrey et al.,

1. In this article, we use the term *implicit* to mean "indirect." Thus, an "implicit measure" assesses mental contents indirectly, in contrast to other forms of measurement that assess mental contents through direct inquiry. Our use of this term contrasts with other definitions that refer to the qualitative nature of the construct (e.g., unconscious) but aligns with the perspective that implicit measures assess evaluations under suboptimal processing conditions (De Houwer & Boddez, 2022). Here, we specify indirect measurement as the defining procedural feature of the task that is the focus of the present research (i.e., the Implicit Association Test; Greenwald et al., 1998).

2. Participants are normatively expected to find compatible trials easier to respond accurately and quickly to. However, we make no assumptions that the stimulus pairings are compatible or incompatible with any given individual's evaluations.

2005), proposes that responses on the IAT reflect not only evaluations of target groups, but also an inhibitory cognitive process that can constrain the expression of evaluations, among other processes. Gonsalkorale et al. (2009) applied the quad model to race IAT responses from older and younger participants to investigate the extent to which evaluations and inhibition contribute to each age group's responses. They found that evaluations were equivalent across age groups, but older people's inhibitory responses were weaker than those of younger people. Thus, the quad model suggests that age-related deficits in inhibition, rather than evaluations, can explain the differences in IAT responses across age groups.

To the extent that a major goal of research using implicit measures is to reduce intergroup biases, MPT models have further advanced theory by providing process-level insight into the mechanisms underlying different bias-reduction interventions. Lai et al. (2014) rank-ordered 18 different interventions configured to reduce implicit racial bias in terms of their effects on *D*-scores. Building on this work, Calanchini et al. (2021) applied the quad model to these same intervention data to develop a process-level intervention taxonomy. These analyses revealed that interventions based on evaluative conditioning selectively targeted control-oriented processes, whereas interventions invoking counterstereotypic exemplars or strategies to override biases affected both evaluative and control-oriented processes. As such, the quad model shores up limitations of *D*-scoring by providing theoretically precise insight into the specific cognitive processes underlying bias-reduction interventions.

That said, the quad model is also limited in some ways. For example, in the context of a race IAT, the quad model assumes that White targets are positively evaluated and Black targets are negatively evaluated. This assumption aligns with how a positive *D*-score is traditionally interpreted to reflect pro-White and anti-Black bias (Greenwald et al., 2003). However, this assumption precludes the possibility that other patterns of evaluations underlie *D*-scores. For example, a positive *D*-score may reflect negative evaluations of Black people and positive evaluations of White people (i.e., the traditional interpretation of *D*-scores). Alternatively, a positive *D*-score could also reflect neutral evaluations of Black people and positive evaluations of White people, or even positive evaluations of Black people and comparatively more positive evaluations of White people. The *D*-scoring algorithm cannot distinguish among these possible patterns of evaluations, and, importantly, neither can the quad model.

To be sure, the White-positive and Black-negative assumption built into the quad model is supported by theories of intergroup relations that articulate why individuals from all segments of society may favor White people and derogate Black people (Clark & Clark, 1950; Jost & Banaji, 1994; Lai & Banaji, 2020; Pratto et al., 2006). For example, system justification theory proposes that people are motivated to defend and justify the legitimacy of the societal status quo (Jost et al., 2004). In Western society, White people are the higher-status group and have greater access to social and material resources than do Black people (Taylor et al., 2011). System justification theory proposes that White people are motivated to defend this system by favoring their racial ingroup and derogating racial outgroups. In addition,

to the extent that Black people are motivated to justify the status quo, they will accept the racial status differential as legitimate and fair by favoring the higher-status racial outgroup and derogating their racial ingroup.

Furthermore, social dominance theory was developed to understand how group-based hierarchies form and are maintained (Pratto et al., 2006). This theoretical perspective proposes a similar route through which people, including subordinated racial groups, might demonstrate pro-White biases. Like system justification theory, social dominance theory posits that some groups (e.g., White people) enjoy greater social status and power than others (e.g., Black people). This group-based hierarchy is maintained through legitimizing myths: consensually held belief systems that provide moral and intellectual justification for group-based discrimination. To the extent that legitimizing myths are agreed upon across all groups, then even members of subordinated groups would be motivated to endorse them, thereby reinforcing their relative position of inferiority within the social hierarchy. Taken together, system justification theory and social dominance theory provide accounts of why both normatively higher- and lower-status racial groups would demonstrate similar patterns of pro-White/anti-Black evaluative preferences, and thus dovetail with the assumptions specified in the quad model.

The quad model's assumptions about White-positive and Black-negative evaluations are especially compelling when considering the race IAT responses of higher-status racial group members (i.e., White people) who are advantaged at multiple levels within systems of inequality, including social institutions and interpersonal interactions. For them, all bases of information should consistently produce positive ingroup and negative outgroup evaluations. However, these assumptions are more difficult to reconcile with evidence that minoritized racial groups sometimes demonstrate ingroup favoritism (Mathur et al., 2010) or ambivalent patterns of implicit bias (Nosek et al., 2002). Instead, we can look to other perspectives on intergroup relations that provide a different framework for understanding bias among members of low-status groups.

The positive-negative asymmetry effect posits that intergroup biases reflect positive evaluations to a greater degree than they do negative evaluations (e.g., Brewer, 1999; Calanchini et al., 2022; Greenwald & Pettigrew, 2014; Halevy et al., 2008; Mummendey & Otten, 1998). For example, intergroup bias may reflect positive evaluations of the ingroup, and less strong but still positive evaluations of the outgroup, such that neither outgroup is associated with negative evaluations. Calanchini et al. (2022) provide empirical evidence for this possibility: they found that Asian and older people's ingroup biases consisted solely of positive evaluations of their lower-status ingroup, absent negative evaluations of the higher-status White or younger outgroup, respectively. Compared to higher-status groups, the biases of low-status group members may be more likely to be driven by positive evaluations of both the ingroup and the outgroup rather, though to different degrees of positivity, than by positive evaluations of the ingroup and negative evaluations of the outgroup, especially given competing motivations to justify the status quo by evaluating higher-status racial outgroup members positively (Jost et al., 2004).

Though racial bias, as measured by the race IAT, could reflect asymmetries in positive evaluations of White versus Black people, we are unable to interrogate this possibility with the quad model as it is traditionally specified. Our goal in the present research is to validate an alternative specification of the quad model configured to estimate positive evaluations of both target racial groups in the race IAT. Specifically, we will test an egalitarian version of the quad model (hereafter referred to as the E-quad model) that assumes IAT responding is driven by positive evaluations of both White and Black people that may nevertheless differ in their levels of positivity.

THE E-QUAD MODEL

Here, we introduce the E-quad model, which draws heavily on the quad model. We begin by describing the theoretical assumptions specified in the quad model and then delineate the major ways in which the E-quad model differs from it. To start, the quad model proposes four qualitatively distinct processes that influence responses on the IAT (Figure 1). The activation of the *Associations*³ parameter refers to the degree to which positive or negative evaluations are activated when responding to a stimulus. All else equal, the stronger the association between the target (e.g., White) and the attribute (e.g., good), the more likely the association is to be activated and drive responses in an evaluation-consistent direction. The quad model estimates two different Associations parameters: one reflects an association between White and good, and the other reflects an association between Black and bad. The *Detection* of correct responses parameter is conceptualized as an accuracy-oriented process and reflects the likelihood that the participant can discern the correct response. Sometimes, activated associations conflict with the detected correct response. For example, on trials in which White faces appear and the categories “White” and “bad” share a response key, to the extent that a participant associates “White” with “good,” then activated associations would conflict with the detected correct response. The quad model proposes an *Overcoming Bias* parameter to resolve such a conflict between Associations and Detection. The Overcoming Bias parameter refers to an inhibitory process that prevents activated associations from influencing responses when they conflict with detected correct responses. Finally, the *Guessing* parameter does not represent a specific process, per se, but instead reflects the tendency to respond with “good” versus “bad” in the absence of influence from the Associations, Detection, and Overcoming Bias parameters. Based on observed patterns of participants’ correct and incorrect responses, the quad model estimates values for the parameters, which are then interpreted as probabilities that each cognitive process influenced participants’ responses on the IAT.

The cognitive architecture of the Quad model is illustrated as a processing tree in Figure 1. The relationships between the parameters are represented through

3. In this article, we retain the associative terminology of the original quad model to describe evaluative mental contents activated by stimuli in the IAT. However, we make no strong assumptions about the representational nature—associative or otherwise—of this construct.

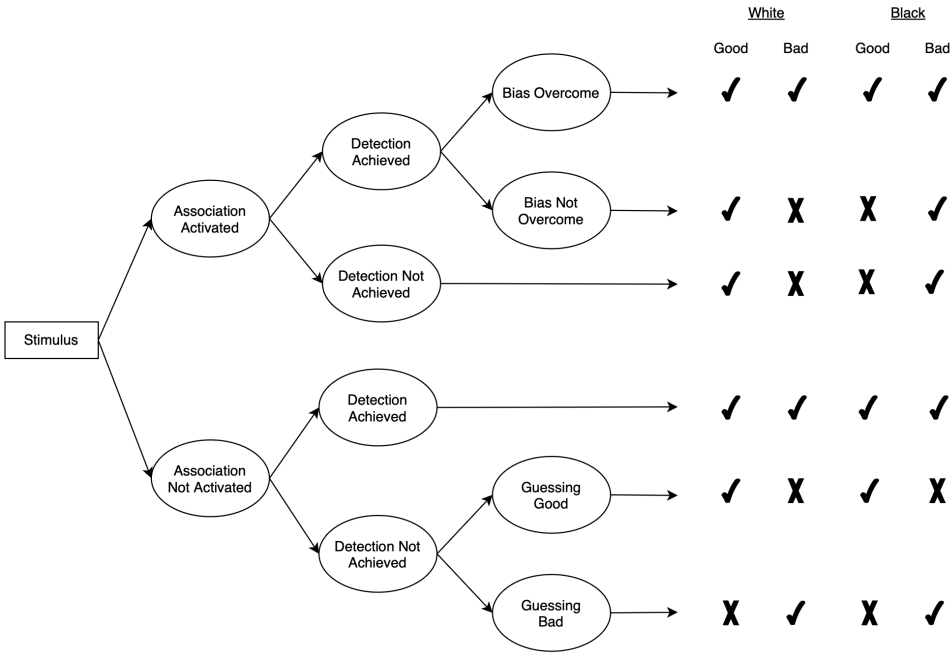


FIGURE 1. Processing tree of the Quad Model.

branching paths, which reflect the sequence of processes that jointly produce correct (✓) and incorrect (X) responses across various trial types of the IAT. Parameters with lines leading to them are conditional upon all preceding parameters. For example, Overcoming Bias is conditional upon the success of activated Associations and Detection.

The E-quad model is illustrated as a processing tree in Figure 2. The primary difference between the quad and E-quad models lies in the assumptions each makes about the directions of evaluation, as reflected in the activated Associations parameters. Both the quad and E-quad models estimate an Associations parameter reflecting White-good associations. However, the quad model also estimates an Associations parameter reflecting Black-bad associations, whereas the E-quad model estimates an Associations parameter reflecting Black-good associations. The diverging implications of these assumptions are reflected in the predicted correct (✓) and incorrect (X) responses in the second and third lines of Figures 1 and 2. In the quad model, activated Black-bad associations will produce an incorrect response on IAT trials in which Black and Good stimuli share a response key, but produce a correct response on trials in which Black and Bad stimuli share a response key. In contrast, in the E-quad model, activated Black-good associations will produce a correct response on IAT trials in which Black and Good stimuli share a response key, but produce an incorrect response on trials in which Black and Bad stimuli share a response key.

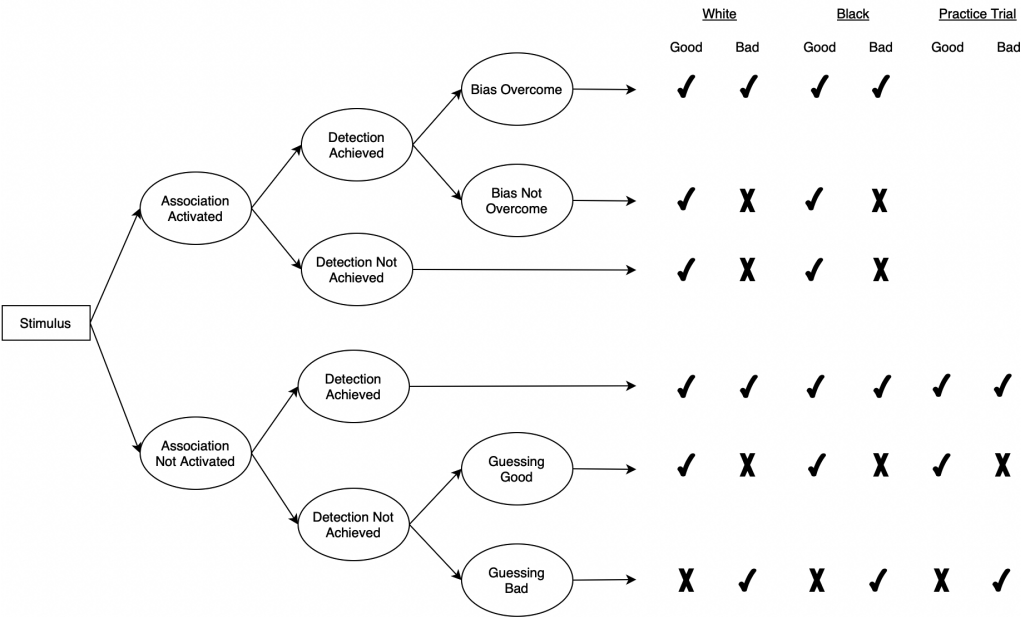


FIGURE 2. Processing tree of the E-quad Model.

The E-quad model also differs from the quad model in terms of the IAT data it uses to estimate its parameters. The quad model relies on data from the so-called critical IAT blocks in which participants respond to both target (i.e., pictures of Black and White people) and attribute (i.e., pleasant and unpleasant words) stimuli. To yield an identifiable model,⁴ the E-quad model also incorporates data from IAT practice blocks in which participants respond only to attribute stimuli. The rightmost column of Figure 2 specifies the role of Detection and Guessing in the IAT on these practice trials, thereby incorporating information on participants’ likelihood of discerning the correct response (i.e., Detection) and their general response tendency towards “good” versus “bad” (i.e., Guessing) free from interfering information related to target (i.e., Black, White) stimuli.

OVERVIEW OF THE E-QUAD MODEL VALIDATION

Thus far, we have proposed a theoretical rationale for the E-quad model and instantiated this theory in the parameters and equations of the MPT model. However,

4. For identified models, the set of predicted response probabilities will only be obtained by a single set of parameter values. The degrees of freedom provided by the number of nonredundant response categories sets an upper limit for the number of identifiable free parameters in a model (Moshagen, 2010) and, generally, a model will not be identifiable if there are more free parameters than non-redundant response categories. In the case of the E-quad model, the Association and Guessing parameters both reflect the tendency to respond with “good,” thereby introducing redundancies in the model equations that are absent in the quad model. As a result, the critical trials of the IAT alone do not provide sufficient degrees of freedom to estimate all parameters of the E-quad model.

neither theory nor equations is sufficient to connect the model parameters to their intended constructs. Instead, such model validation consists of three main steps: we must demonstrate that the model sufficiently accounts for observed patterns of IAT responses, that the model parameters can vary independently of one another, and that the parameters respond in theoretically predicted ways to selective influence (Hütter & Klauer, 2016). We attempted to do so here over the course of six experiments. In all six experiments, we quantified the extent to which the model accounts for observed responses using established fit indices, and five of these experiments introduced selective-influence manipulations intended to facilitate or constrain specific cognitive processes related to Associations, Detection, Overcoming Bias, and Guessing. We can infer that a parameter reflects its intended construct if the manipulation affects the relevant parameter in the predicted direction (i.e., convergent validity) and does not affect any of the irrelevant parameters (i.e., discriminant validity).

The first experiment consisted of two separate race IAT measurement occasions administered longitudinally without any manipulations, which provided us with an initial opportunity to assess model fit twice and to investigate the temporal stability of E-quad model parameters. The second experiment sought to show that Detection and Overcoming Bias are constrained by a short IAT response window. The third experiment investigated whether Guessing varies according to the base rates of good and bad attribute stimuli. The fourth experiment explored how the Association parameters respond to exposure to counterstereotypical admired Black exemplars and reviled White exemplars. The fifth and sixth experiments relied on a known-groups paradigm and were designed to show that Associations are sensitive to participant identity in the contexts of race and sexual orientation, respectively.⁵

Additionally, given that the E-quad model is closely related to the quad model, we compared the two models using selection indices following the selective-influence experiments to determine which model better accounts for observed IAT responses. To this end, we reported two Bayesian-oriented model selection indices: the deviance information criterion (DIC: Klauer, 2010; Spiegelhalter et al., 2002) and the widely applicable information criterion (WAIC: Vehtari et al., 2017; Watanabe, 2010). Given the theoretical rationale we articulated above, we expected the E-quad model to show better fit than the quad model, especially for data from members of lower-status groups compared to higher-status groups.

THE PRESENT RESEARCH

The present research was approved by the Institutional Review Board of the University of California, Riverside #HS 20-278. For all lab-based experiments (i.e., Experiments 1–4), we aimed to recruit approximately 100 participants per condition, which is approximately double the average published sample size in similar

5. We conducted one additional experiment to test the effects of cognitive load on model parameters, but the manipulation failed. Nevertheless, for research transparency, we report the results in Tables S7 and S8 in the Supplemental Materials.

modeling research. Unless otherwise noted, all hypotheses and analyses were pre-registered. Preregistrations, code, data, and supplemental materials are available at https://osf.io/5y7x3/?view_only=fd5ee3e895314e6a8595e9a36a8f24ac.

MEASURES AND ANALYTIC PLAN

All experiments rely on the IAT and MPT models as an alternative IAT scoring approach. We describe these methodological details below.

Race Implicit Association Test. The majority of our experiments relied on the race IAT, which consists of stimuli reflecting two target categories (Black, White) and two attribute categories (good, bad). The IAT proceeds in seven blocks. The first block consists of 20 practice trials in which participants categorize pictures of Black and White people. The second block consists of 20 practice trials in which participants categorize words as good or bad. In the third and fourth critical blocks, the stimuli and response keys are combined so that participants complete a total of 60 trials in which they respond to White and good stimuli with one response key and to Black and bad stimuli with the other. The fifth block consists only of Black and White stimuli, and participants complete 20 practice trials with the response mapping reversed relative to the mapping in the previous blocks. In the sixth and seventh critical blocks, participants completed a total of 60 trials in which they responded to White and bad stimuli with one response key and to Black and good stimuli with the other.

Multinomial Processing Tree Models. Multinomial processing tree (MPT; Batchelder & Riefer, 1999) models are a class of formal mathematical models that link latent processes to observable responses on tasks such as the IAT. MPT models are tailored to specific experimental paradigms that provide frequency data (e.g., number of correct and incorrect responses). They serve as an alternative to the *D*-scoring algorithm for quantifying IAT performance by specifying the number, nature, and composition of cognitive processes thought to contribute to task responses (Hütter & Klauer, 2016).

To quantify the influence of each process specified in each MPT model, we implemented a Bayesian estimation approach for hierarchical latent-trait MPT models (Klauer, 2010). We fitted all hierarchical MPT models using default priors in *TreeBUGS* (Heck et al., 2018) in R Programming Environment v.4.2.2, which draws posterior samples of the parameters using Markov chain Monte Carlo (MCMC) methods. We used a conservative estimation approach with 3 chains, each with 100,000 iterations, discarding the first 10,000 as the burn-in period. To ensure sufficient parameter convergence across all models, we relied on a criterion of a Gelman-Rubin $\hat{R} < 1.05$ and a visual inspection of Gelman-Rubin trace plots.

All parameters are estimated on a probit scale and then transformed to a probability scale, ranging from 0 to 1. Higher values on the Association, Detection, and Overcoming Bias parameters can be interpreted to reflect greater influence of each cognitive process. The Guessing parameter was coded so that values greater than

.5 reflect a bias toward responding with the “good” key, and values less than .5 reflect a bias toward responding with the “bad” key. For each experiment, we estimated a single MPT model that included the experimental condition as a discrete fixed-effects covariate, thereby producing separate parameter estimates for each experimental group.

EXPERIMENT 1

The primary goal of Experiment 1 was to apply the E-quad model to the race IAT. We estimated two Associations parameters, one intended to reflect positive evaluations of White targets and another intended to reflect positive evaluations of Black targets. We estimated a single Detection parameter that reflects the ability to determine the correct response across all stimuli presented in the IAT. We estimated an Overcoming Bias parameter, which reflects inhibiting Associations on trials in which they conflict with a successfully detected correct response (i.e., incompatible trials). In contrast to both the quad model and the *D*-scoring algorithm, here we define incompatible trials in the E-quad model as trials in which “bad” stimuli share a response key with either Black or White targets. Lastly, we estimated a single Guessing parameter, which reflects a general response bias toward “good” versus “bad” in the absence of influence from the other processes specified in the E-quad model. Given that this was our first application of the E-quad model to IAT data, our plan was to evaluate model fit and the theoretical interpretability of the parameter estimates. Additionally, in Experiment 1, we administered the IAT to each participant on two measurement occasions, enabling us to examine the stability of the parameter estimates over time.

Participants. For partial course credit, 132 undergraduate students participated in the experiment. Participants completed two online assessments administered longitudinally, with a 24–48-hour window between the first and second sessions. In each session, participants completed a race IAT followed by demographic measures. Of these participants, 103 (62% women, 35% men, 3% nonbinary or gender nonconforming) were included in the analysis after excluding those who did not fully complete the IAT at both measurement occasions. 50% of participants identified their race or ethnicity as Asian, 30% as Hispanic or Latino, 8% as White, 4% as Black or African American, 2% as Middle Eastern or North African, 1% as American Indian or Alaska Native, 4% as multiracial, and 1% who preferred not to answer.

Results. To evaluate model fit, we calculated w for the T_1 statistic (Klauer, 2010). The T_1 statistic summarizes how well the model accounts for the pattern of observed response frequencies (i.e., the number of correct and incorrect responses on the IAT) aggregated across participants and corresponds to the goodness-of-fit chi-square statistic used in traditional modeling approaches (Batchelder & Riefer, 1999). As such, T_1 depends on the number of observations, with more observations providing greater power to detect even a small degree of model misfit. Consequently, we

rely on w , which scales T_1 by the number of observations (Equation 1). Based on Cohen's (1992) effect size recommendations, we interpret a $w < 0.1$ as indicating acceptable model fit in the present research. Using this approach, we found that the E-quad model provided an acceptable fit to these data ($w = 0.009$).

$$W = \sqrt{\frac{T_{\text{observed}}}{n \text{ participants} * n \text{ trials}}} \quad (1)$$

Parameter estimates for the E-quad model were generally interpretable (Table 1). The 95% Bayesian credible intervals (BCIs) for both the Associations-White and Associations-Black parameter estimates excluded zero, which indicates that they credibly influenced responses. At the same time, neither Associations parameter estimate exceeded 0.05, suggesting that the likelihood that racial associations drive IAT responses was small. The Detection estimates of .92 and .90 were relatively high, suggesting that Detection plays a prominent role in IAT responding. 95% BCIs for Overcoming Bias estimates at both measurement occasions spanned a large portion of the [0,1] likelihood scale, making these parameters difficult to interpret. Guessing did not differ from .50, suggesting no general response bias toward "good" versus "bad" stimuli.

We used planned contrasts to explore differences between the Associations-White and Associations-Black parameter estimates. We carried out this analysis by subtracting the distributions for all MCMC posterior samples of the Associations-Black parameter from the distributions for all MCMC posterior samples of the Associations-White parameter at each time point. In the resulting distribution of credible mean differences, differences between parameter estimates are credible if the 95% BCI does not contain zero. The results reveal that the Associations-White parameter was higher than the Associations-Black parameter at time 1 (mean difference = .04, BCI95% [.02, .07]), but not at time 2 (mean difference = .01, BCI95% [-.007, .03]).

We also used planned contrasts to assess the extent of mean-level parameter stability in IAT responding from time 1 to time 2 (Table 1). For each model parameter, we subtracted the MCMC posterior samples of time 2 responses from the MCMC posterior samples of time 1 responses. Of the five processes specified in the E-quad model, we identified reliable differences between time points in Associations-White and Detection, with both parameters higher at time 1 than at time 2. Overcoming Bias and Guessing did not differ reliably across measurement occasions. In addition, we explored rank-order stability by correlating individual-level parameter estimates at time 1 and time 2 for all E-quad parameters. We found a moderately strong, positive correlation between time 1 and time 2 parameter estimates for Detection. However, we did not find reliable correlations between time 1 and time 2 parameter estimates for Associations-White, Associations-Black, Overcoming Bias, and Guessing.

Discussion. In Experiment 1, the E-quad model provided an acceptable fit to the observed pattern of IAT responses. Additionally, the E-quad model produced theoretically interpretable parameter estimates. White-positive associations were

TABLE 1. Parameter Estimates, Planned Contrast, Time 1-Time 2 Correlations, and 95% BCI for Experiment 1

Process	Time 1 Mean Estimate	Time 2 Mean Estimate	Mean Differences	Time 1 - Time 2 Correlation
Associations-White	.05 [.03, .07]	.02 [.002, .04]	.04 [.009, .06]	.46 [−.14, .83]
Associations-Black	.009 [< .001, .03]	.006 [< .001, .02]	.003 [−.01, .02]	.07 [−.67, .70]
Detection	.92 [.91, .93]	.90 [.87, .92]	.02 [.002, .04]	.68 [.49, .82]
Overcoming Bias	.89 [.58, .99]	.61 [.07, .99]	.28 [−.20, .85]	.05 [−.73, .80]
Guessing	.46 [.40, .51]	.49 [.43, .54]	−.03 [−.11, .05]	.08 [−.71, .77]

Note. Brackets indicate 95% Bayesian credible intervals. Bold reflects credible mean differences and correlations.

reliably higher than Black-positive associations, dovetailing with research indicating that higher-status groups are generally evaluated more positively than lower-status groups (Nosek et al., 2007), even among a participant sample like ours that primarily consists of members of lower-status groups. Detection parameter estimates were generally very high, which corresponds with the high degree of accuracy participants displayed on the IATs in Experiment 1 (6% and 7% error rate at time 1 and 2, respectively). Overcoming Bias parameters were estimated with a high degree of imprecision, which makes them difficult to interpret. However, the same is generally true of the Overcoming Bias parameter as specified in the quad model for individual-level analyses (Elder et al., 2023).

This first experiment demonstrated that the E-quad model can fit data from a race IAT, a necessary first step toward validating a new model specification. To further support the construct validity of the E-quad model, we must also demonstrate that each parameter is sensitive to selective influence in accordance with established theory. Specifically, target parameters should respond to experimental manipulations (i.e., convergent validity) without affecting unrelated parameters (i.e., discriminant validity). In Experiments 2 through 6, we conducted a series of selective influence experiments to examine whether the parameters of the E-quad model vary meaningfully and independently of one another.

EXPERIMENT 2

The aim of Experiment 2 was to show that processes conceptualized as control-oriented will be constrained by limited time to respond to the IAT. Considerable previous research has demonstrated that control-oriented processes require more time to influence responses than do processes conceptualized as automatic

(for a review, see Moors & De Houwer, 2006). Detection is conceptualized as an accuracy-oriented process, and Overcoming Bias is conceptualized as an inhibitory process, and, as such, both exert control over response tendencies. Consequently, we expect that imposing a short response window of 750ms on the IAT will decrease the likelihood that Detection and/or Overcoming Bias contribute to participants' responses (Cunningham et al., 2001). However, the Associations and Guessing parameters are conceptualized to reflect automatic processes and, thus, should not be affected by a short response window.

Participants and Method. For partial course credit, 199 undergraduate students participated in the experiment. Of these participants, 181 (66% women, 30% men, and 4% nonbinary, gender nonconforming, or prefer not to answer) were included in the analysis after removing participants who failed the attention check, attempted to take the experiment twice, or requested their data not be analyzed. Participants were randomly assigned to complete a standard race IAT ($n = 87$) or a race IAT with a 750ms response window ($n = 94$). After taking the IAT, participants completed demographic questionnaires. 49% of participants identified their race or ethnicity as Asian, 30% as Hispanic or Latino, 4% as White, 4% as Black or African American, 5% as Middle Eastern or North African, 4% as multiracial, .5% as Native Hawaiian or other Pacific Islander, and 3.5% who preferred not to answer.

Results. The E-quad model provided an acceptable fit to these data ($w = 0.009$). We used planned contrast to explore process-level differences between the standard IAT procedure and the IAT with a short response window. For each model parameter, we subtracted the MCMC posterior samples of the response window condition from the MCMC posterior samples of the standard IAT condition.

In Experiment 2, the E-quad model produced parameter estimates that were mostly interpretable. Both Associations-White and Associations-Black parameter estimates were relatively small, and Associations-White estimates were higher than Associations-Black estimates across each experimental condition. Detection estimates were relatively high in the standard IAT condition, though less so in the response window condition. Mirroring the results of Experiment 1, Overcoming Bias parameters were estimated with very wide BCIs. Guessing estimates did not credibly differ from .5 in either the standard IAT condition or in the response window condition. Table 2 summarizes the parameter values and planned contrasts.

We used a planned contrast to examine process-level differences between IAT responses in the response-window condition and those in the standard IAT condition. For each model parameter, we subtracted the MCMC posterior samples of standard IAT responses from the MCMC posterior samples of the response window responses. Of the processes specified in the E-quad model, we identified reliable differences between experimental conditions in Associations-White, Associations-Black, and Detection. Contrary to predictions, Associations-White was higher in the response-window condition than in the standard IAT condition. Associations-Black was also higher in the response window condition compared to the standard IAT condition. As expected, Detection was lower in the response

TABLE 2. Parameter Estimates, Planned Contrast, and 95% BCI for Experiment 2

Process	Short Response Window: Mean Estimate	Standard Procedure: Mean Estimate	Mean Difference
Associations-White	.14 [.09, .19]	.04 [.02, .06]	.10 [.05, .15]
Associations-Black	.03 [.002, .09]	.003 [< .001, .01]	.02 [< .001, .09]
Detection	.77 [.73, .80]	.90 [.88, .92]	-.14 [-.17, -.10]
Overcoming Bias	.77 [.13, .99]	.62 [.01, .99]	.14 [-.58, .96]
Guessing	.44 [.39, .49]	.46 [.41, .51]	-.02 [-.09, .05]

Note. Brackets indicate 95% Bayesian credible intervals. Bold reflects credible mean differences.

window condition than in the standard IAT condition. Neither Guessing nor Overcoming Bias varied as a function of experimental condition.

Discussion. The results of Experiment 2 offer mixed support for the construct validity of the E-quad model. We predicted and found that a short response window decreased Detection estimates, which supports our conceptualization of Detection as a control-oriented process (Conrey et al., 2005). We predicted and found that a short response window decreased Detection estimates, which supports our conceptualization of Detection as a control-oriented process (Conrey et al., 2005). As predicted, we found no differences in Guessing as a function of experimental condition. However, we did not find differences in Overcoming Bias across conditions. Though we expected Overcoming Bias to be sensitive to a short response window, this null result aligns with past research using the Quad model, suggesting that low precision in Overcoming Bias estimates limits statistical inference, especially in smaller sample sizes (Elder et al., 2023). Additionally, the short response window unexpectedly influenced both Associations parameters, which is not consistent with past interpretations of Associations as an automatic process (Beer et al., 2008; Conrey et al., 2005). Specifically, Associations-White and Associations-Black were higher in the response-window condition when controlled processing was constrained, than in the standard IAT condition.

EXPERIMENT 3

The main goal of Experiment 3 was to provide validation evidence for the Guessing parameter. Guessing is operationalized to reflect a bias towards “good” or “bad” responses in the absence of influence from other cognitive processes. In Experiment 3, we manipulated the number of good and bad attribute trials in the

race IAT and expected that the Guessing parameter would be selectively influenced, but other processes would remain unaffected. Given that Guessing parameter values greater than .5 indicate a positive response bias and values less than .5 indicate a negative response bias, we predicted that Guessing parameter estimates among participants who see more good attributes would be higher than .5 whereas Guessing parameter estimates among participants who see more bad attributes would be lower than .5.

Participants and Method. For partial course credit, 213 undergraduate students participated in the experiment. Of these participants, 194 (73% women, 25% men, and 2% nonbinary or gender nonconforming) were included in the analysis after removing participants who failed the attention check, attempted to take the experiment twice, or requested their data not be analyzed. In a between-subjects design, participants were randomly assigned to the good or bad base-rate condition and completed demographic questionnaires after completing the IAT. A total of 95 participants completed a race IAT with 75% good attributes, and 99 participants completed a race IAT with 75% bad attributes. 34% of participants identified their race or ethnicity as Asian, 36% as Hispanic or Latino, 10% as White, 3% as Black or African American, 8% as Middle Eastern or North African, 7% as multiracial, .5% as American Indian or Alaska Native, and 1.5% who preferred not to answer.

Results. The E-quad model demonstrated an acceptable fit ($w = 0.006$), and the parameter estimates mirrored those from previous experiments. Replicating patterns from Experiments 1 and 2, Associations-White estimates were higher than Associations-Black estimates across both base-rate conditions. Detection estimates remained relatively high, and Overcoming Bias had BCIs that spanned nearly the entire range of possible values.

We used a planned contrast to explore process-level differences in IAT responses between the positive and negative base-rate conditions. For each model parameter, we subtracted the MCMC posterior samples of positive base-rate responses from the MCMC posterior samples of negative base-rate responses. Of the processes specified in the E-quad model, we identified credible differences in all processes except Detection. Associations-White and Associations-Black were higher in the positive base-rate condition than in the negative base-rate condition. Overcoming Bias was higher in the positive base-rate condition than in the negative base-rate condition. As expected, Guessing was higher in the positive base-rate condition than in the negative base-rate condition. In addition, Guessing was credibly lower than .50 in the negative base-rate condition and credibly higher than .50 in the positive base-rate condition. Table 3 summarizes the parameter values and planned contrasts.

Discussion. The results of Experiment 3 offer mixed support for the construct validity of the E-quad model. We predicted and identified an effect of base-rate asymmetry on Guessing, which has previously been conceptualized as a general response bias toward good or bad in the absence of other cognitive influences

TABLE 3. Parameter Estimates, Planned Contrast, and 95% BCI for Experiment 3

Process	Negative Base Rate: Mean Estimate	Positive Base Rate: Mean Estimate	Mean Difference
Association-White	0.03 [< .001, 0.12]	0.09 [< .001, 0.13]	−0.06 [−0.10, −0.02]
Association-Black	0.002 [< .001, 0.006]	0.02 [0.0016, 0.07]	−0.02 [−0.08, −0.001]
Detection	0.89 [0.72, 0.79]	0.91 [0.87, 0.91]	−0.02 [−0.05, 0.01]
Overcoming Bias	0.01 [< .001, 0.94]	0.98 [0.034, 1.000]	−0.97 [−1.00, −0.68]
Guessing	0.22 [0.17, 0.26]	0.62 [0.56, 0.67]	−0.40 [−0.46, −0.33]

Note. Brackets indicate 95% Bayesian credible intervals. Bold reflects credible mean differences.

(Conrey et al., 2005). Experimentally varying the base rate at which participants see good versus bad attributes on the IAT should influence general response tendencies, as assessed by the Guessing parameter. Accordingly, we saw that participants assigned to the positive base rate condition had higher Guessing parameter estimates compared to participants assigned to the negative base rate condition. However, the base-rate asymmetry unexpectedly influenced both Associations parameters and Overcoming Bias, thus failing to provide evidence of discriminant validity.

EXPERIMENT 4

The main goal of Experiment 4 was to provide validation evidence for the Associations parameters. Prior work has suggested that exposure to counterstereotypic exemplars, such as admired Black and reviled White exemplars, reduces pro-White/anti-Black evaluations (Calanchini et al., 2021; Dasgupta & Greenwald, 2001). Consequently, we predicted that replacing the stimuli depicting unknown Black and White male faces that are typically used in the IAT with faces of famous, admired Black individuals (e.g., Martin Luther King, Jr.) and infamous, reviled White individuals (e.g., Adolf Hitler) should reduce the magnitude of preference for White people over Black people. If Associations-White and Associations-Black reflect the psychological processes we propose, then Associations-White parameter estimates should be lower among participants who see famous negative White exemplars on the IAT compared to participants who see unknown target stimuli on a standard IAT. Similarly, Associations-Black parameter estimates should be higher among participants who see famous positive Black exemplars than among participants who see unknown target stimuli.

Participants and Method. For partial course credit, 187 undergraduate students participated in the experiment. Of these participants, 180 (74% women, 23% men, and 3% non-binary) were included in the analysis after excluding those who failed the attention check, attempted to take the experiment twice, or requested that their data not be analyzed. In a between-subjects design, participants were randomly assigned to complete a standard race IAT ($n = 90$) or a race IAT with counterstereotypic exemplars ($n = 90$). In the counterstereotypic exemplars condition, participants first viewed the faces of each admired Black and reviled White target, along with a short description of each, before beginning the IAT. In the standard race IAT condition, participants viewed the unknown Black and White targets before beginning the IAT. After taking the IAT, participants completed demographic questionnaires. 43% of participants identified their race or ethnicity as Asian, 30% as Hispanic or Latino, 4% as White, 3% as Black or African American, 8% as Middle Eastern or North African, 11% as multiracial, and 1% who preferred not to answer.

Materials. We selected the counterstereotypic exemplars based on pretesting. To pretest the stimuli, we presented pictures of ten Black individuals we assumed were admired and ten White individuals we assumed were reviled to a separate group of 192 participants and asked them to rate each exemplar on identifiability and likeability. After averaging identifiability scores and likability scores across each exemplar, we retained the four most identifiable and likable Black individuals (Martin Luther King, Jr., Morgan Freeman, Kobe Bryant, Chadwick Boseman) and the four most identifiable and unlikable White individuals (Ted Bundy, Adolf Hitler, Derek Chauvin, Donald Trump). Importantly, the participant pool we relied on for stimuli pretesting was the same as we used for Experiment 4.

Results. The E-quad model demonstrated an acceptable fit ($w = 0.006$), and parameter estimates generally fell within interpretable ranges. Interestingly, Guessing parameter estimates were reliably lower than .5 in both experimental conditions, suggesting a response bias towards “bad” attributes.

We used planned contrast to explore process-level differences between IAT responding in the counterstereotypic exemplar condition versus the standard IAT condition. For each model parameter, we subtracted the MCMC posterior samples of standard IAT responses from the MCMC posterior samples of counterstereotypic exemplar responses. However, we did not identify any reliable differences in E-quad parameters between experimental conditions. Table 4 summarizes the parameter values and planned contrasts.

Discussion. The results of Experiment 4 failed to provide construct validity for the Associations parameters of the E-quad model. We predicted that participants who saw counterstereotypic exemplars on an IAT would have lower Associations-White parameter estimates, indicating less positive evaluations, and higher Associations-Black parameter estimates, indicating more positive evaluations, compared to participants who completed a standard race IAT. The expected effects did not emerge, which is surprising given that previous research has successfully

Table 4. Parameter Estimates, Planned Contrast, and 95% BCI for Experiment 4

Process	Exemplar IAT: Mean Estimate	Standard IAT: Mean Estimate	Mean Difference
Association-White	.04 [.02, .06]	.05 [.03, .08]	–.01 [–.04, .01]
Association-Black	.007 [< .001, .03]	.003 [< .001, .01]	0.004 [–.01, .02]
Detection	.90 [.88, .92]	.92 [.90, .94]	–0.02 [–.04, .01]
Overcoming Bias	.37 [< .001, .97]	.45 [.01, 0.99]	–.08 [–.93, .79]
Guessing	.43 [.36, .49]	.42 [.35, .48]	0.01 [–.08, .10]

Note. Brackets indicate 95% Bayesian credible intervals. Bold reflects credible mean differences.

used the same paradigm to demonstrate the malleability of responses on the IAT (Dasgupta & Greenwald, 2001; Gonsalkorale et al., 2010). That said, recent research has also failed to replicate the effects of counterstereotypic exemplars on IAT responses (Kurdi et al., 2024). In light of these challenges, we carried out two additional experiments to explore whether the Associations parameters could be selectively influenced.

EXPERIMENT 5A

One of the most replicable findings in intergroup relations research is preferential evaluations of ingroup members relative to outgroup members (Brewer, 1979; Tajfel, 1982; Tajfel et al., 1971). In the context of implicit social cognition, race IAT scores consistently vary with the respondent’s racial identity. Specifically, Black participants show significantly less pro-White bias on a Black-White race IAT compared to White participants (Blair et al., 2015; Gonsalkorale et al., 2010). Drawing from this past research, we carried out a known-groups experiment to test whether Associations parameters covaried with respondents’ racial identity. Using a Black-White race IAT, we predicted that White respondents would have higher Associations-White parameter estimates compared to Black respondents. Likewise, we predicted that Black respondents would have higher Associations-Black parameter estimates compared to White respondents.

Participants. Experiment 5a relied on open-source data freely available from Project Implicit (implicit.harvard.edu). From the year 2020, we randomly selected 100 Project Implicit visitors who self-identified as Black and 100 who self-identified as White. All visitors completed a race IAT.

Results. The E-quad model demonstrated acceptable fit ($w = 0.005$), and the parameter estimates tended to fall within interpretable ranges. We used planned contrast to explore process-level differences between White and Black IAT respondents. For each model parameter, we subtracted the MCMC posterior samples of White participants' IAT responses from the MCMC posterior samples of Black participants' IAT responses. Though we did not identify any reliable differences in E-quad parameters between participant groups in planned contrasts, we did find an exploratory contrast that Associations-White was higher than Associations-Black among White respondents (mean difference = .03, BCI95% [.01, .05]). Table 5 summarizes the parameter values and planned contrasts.

Discussion. The results of Experiment 5a failed to provide evidence of construct validity for the Associations parameters of the E-quad model, which did not vary between White and Black participants, as we predicted. In all experiments thus far, we have relied on 100 participants per condition as a heuristic. However, in the context of MPT modeling, statistical power depends on both sample size and the precision with which a given parameter is estimated. To investigate whether 100 participants per condition were insufficient in detecting the effects of the Associations parameters, we conducted a well-powered follow-up study with 10,000 participants per condition.

EXPERIMENT 5B

Participants. Experiment 5b also relied on open-source data freely available from Project Implicit (implicit.harvard.edu). From the year 2020, we randomly selected 10,000 Project Implicit visitors who self-identified as Black and 10,000 who self-identified as White. All visitors completed a race IAT.

TABLE 5. Parameter Estimates, Planned Contrast, and 95 % BCI for Experiment 5a

Process	White Respondents: Mean Estimate	Black Respondents: Mean Estimate	Mean Difference
Association-White	.03 [.01, .05]	.02 [.009, .04]	-.01 [-.03, 0.01]
Association-Black	.003 [< .001, .01]	.01 [< .001, .03]	.01 [-.002, .03]
Detection	.92 [.90, .94]	.91 [.89, .93]	-.01 [-0.03, .02]
Overcoming Bias	.66 [< .001, > .999]	.17 [.04, .81]	-.49 [-0.99, 0.25]
Guessing	.43 [.37, .49]	.44 [.38, .50]	.01 [-.07, .09]

Note. Brackets indicate 95% Bayesian credible intervals. Bold reflects credible mean differences.

Results. The E-quad model demonstrated acceptable fit ($w = 0.0005$), and the parameter estimates tended to fall within interpretable ranges. We used planned contrast to explore process-level differences between White and Black IAT respondents. For each model parameter, we subtracted the MCMC posterior samples of White participants' IAT responses from the MCMC posterior samples of Black participants' IAT responses. As expected, Associations-White was higher among White respondents compared to Black respondents. In addition, Associations-Black was higher among Black respondents compared to White respondents. Detection was higher among White respondents than among Black respondents, and Guessing was higher among Black respondents than among White respondents. Table 6 summarizes the parameter values and planned contrasts.

Discussion. Experiment 5b provides partial support for the predicted relationship between participants' identity and their evaluations. As a result of the greater statistical power conferred by increased sample size, we identified reliable racial differences in four quad parameters. As predicted, White participants demonstrated reliably higher positive evaluations of White participants than did Black participants, and Black participants demonstrated reliably higher positive evaluations of Black people than did White participants. Detection and Guessing parameters also varied by race. Though we have no theory to support a prediction that either of these parameters should vary according to participant race, Gonsalkorale et al. (2010) also found that Detection was lower for Black than White Project Implicit visitors.

EXPERIMENT 6A

In a further attempt to selectively manipulate the Associations parameters, we conducted another known-groups experiment on sexual orientation. The procedural details of Experiment 6a were similar to those of experiments 5a and 5b. We carried out a known-groups experiment to test whether Associations parameters covaried with respondents' sexual orientation. Using a sexuality IAT, we predicted that straight respondents would have higher Associations-Straight parameter estimates compared to gay respondents. Likewise, we predicted that gay respondents would have higher Associations-Gay parameter estimates compared to straight respondents.

Participants. Experiment 6a also relied on open-source data freely available from Project Implicit (implicit.harvard.edu). From the year 2020, we randomly selected 100 Project Implicit visitors who self-identified as gay and 100 who self-identified as straight. All visitors completed a sexuality IAT.

Results. The E-quad model demonstrated acceptable fit ($w = 0.005$). We used planned contrast to explore process-level differences between gay and straight IAT respondents. For each model parameter, we subtracted the MCMC posterior samples of gay participants' IAT responses from the MCMC posterior samples of straight participants' IAT responses. Of the processes specified in the E-quad model, we identified reliable differences in Associations-Straight and Guessing.

TABLE 6. Parameter Estimates, Planned Contrast, and 95% BCI for Experiment 5b

Process	White Respondents: Mean Estimate	Black Respondents: Mean Estimate	Mean Difference
Association-White	.042 [.04, .045]	.022 [.020, .024]	.02 [.016, .023]
Association-Black	.0004 [< .001, .001]	.003 [.002, .005]	-.003 [-.005, -.002]
Detection	.92 [.92, .93]	.91 [.90, .91]	.02 [.015, .023]
Overcoming Bias	< .001 [< .001, .004]	.003 [< .001, .03]	-.003 [-0.027, .001]
Guessing	.39 [.38, .39]	.43 [.43, .44]	-.048 [-.059, -.037]

Note. Brackets indicate 95% Bayesian credible intervals. Bold reflects credible mean differences.

As expected, Associations-Straight was higher among straight respondents compared to gay respondents. Unexpectedly, Guessing was higher among gay respondents compared to the straight respondents, suggesting that gay respondents have a relatively more positive general response bias. In an exploratory analysis, we found that Associations-Straight was higher than Associations-Gay for straight respondents (mean difference = .045, BCI95% [.002, .15]) but not for gay respondents (mean difference = -.01, BCI95% [-.06, .01]). Table 7 summarizes the parameter values and planned contrasts.

Discussion. Experiment 6a also provides partial support for the predicted relationship between participants’ identity and their evaluations. Straight participants demonstrated reliably higher positive evaluations of straight people than did gay participants. However, gay evaluations did not reliably differ between straight and gay participants. Mirroring Experiment 5b, we conducted another well-powered follow-up study that included 10,000 participants per condition.

EXPERIMENT 6B

Participants. Experiment 6b also relied on open-source data freely available from Project Implicit (implicit.harvard.edu). From the year 2020, we randomly selected 10,000 Project Implicit visitors who self-identified as gay and 10,000 who self-identified as straight. All Project Implicit visitors completed a sexuality IAT.

Results. The E-quad model demonstrated acceptable fit ($w = 0.0005$). We used planned contrast to explore process-level differences between gay and straight IAT respondents. For each model parameter, we subtracted the MCMC posterior

TABLE 7. Parameter Estimates, Planned Contrast, and 95% BCI for Experiment 6a

Process	Straight Respondents: Mean Estimate	Gay Respondents: Mean Estimate	Mean Difference
Association-Straight	.05 [.004, .16]	.004 [< .001, .01]	.04 [.003, .15]
Association-Gay	.003 [< .001, .01]	.01 [< .001, .06]	-.01 [-.06, .005]
Detection	.89 [.86, .92]	.86 [.83, .89]	.03 [-.01, .07]
Overcoming Bias	.10 [< .001, .73]	.25 [< .001, > .999]	-.15 [< -.999, 0.43]
Guessing	.44 [.37, .50]	.53 [.47, .59]	-.09 [-.18, -.01]

Note. Brackets indicate 95% Bayesian credible intervals. Bold reflects credible mean differences.

samples of gay participants’ IAT responses from the MCMC posterior samples of straight participants’ IAT responses. Of the processes specified in the E-quad model, we identified reliable differences in Associations-Straight, Associations-Gay, and Guessing.

As expected, Associations-Straight was higher among straight respondents compared to gay respondents. Likewise, Associations-Gay was higher among gay respondents compared to straight respondents. Guessing was higher among gay respondents compared to the straight respondents, suggesting that gay respondents have a relatively more positive general response bias. Table 8 summarizes the parameter values and planned contrasts.

Discussion. Experiment 6b provides support for the predicted relationship between participants’ identity and their evaluations. Straight participants demonstrated reliably higher positive evaluations of straight people than did gay participants. In addition, gay participants demonstrated reliably higher positive evaluations of gay people than did straight people.

COMPARING THE E-QUAD MODEL WITH THE QUAD MODEL

The results of Experiments 1–6 do not provide compelling evidence for the validity of the E-quad model. Though the E-quad model showed adequate fit to IAT data across all experiments, the model’s parameters did not consistently respond to experimental manipulations in theoretically predicted directions. Specifically, the Detection, Guessing, and Associations parameters were (generally) affected by experimental manipulations targeted for each of them. However, other parameters beyond the targeted ones were also affected by these manipulations. That said,

TABLE 8. Parameter Estimates, Planned Contrast, and 95% BCI for Experiment 6b

Process	Straight: Mean Estimate	Gay: Mean Estimate	Mean Difference
Association-Straight	.07 [.01, .19]	.008 [< .001, .03]	.06 [.004, .1]
Association-Gay	.001 [< .001, .003]	.01 [< .001, .02]	-.01 [-.01, -.0004]
Detection	.88 [.86, .89]	.88 [.87, .90]	-.005 [-.02, .01]
Overcoming Bias	.01 [< .001, .09]	.74 [< .001, 1]	-.73 [-1, < .001]
Guessing	.41 [.35, .47]	.49 [.43, .55]	-.09 [-.14, -.03]

Note. Brackets indicate 95% Bayesian credible intervals. Bold reflects credible mean differences.

the E-quad model aligns very closely with the quad model (Conrey et al., 2005) that inspired it. Consequently, we directly compared the E-quad and quad models using data from Experiments 1–6.

We applied the quad model to the same data reported for Experiments 1–6 and then conducted planned contrasts using the same method described above for the E-quad model. We report the quad parameter values and planned contrasts in full in Tables S1–S6 of the supplemental materials and summarize our findings here.

Our findings when applying the quad model (Conrey et al., 2005) to the IAT data are as follows. Across all experiments, the quad model provided an acceptable fit to IAT data. The goal of Experiment 1 was to collect race IAT data and evaluate the interpretability of the quad model parameters. Quad model estimates for Experiment 1 data fall within interpretable ranges, with the control-oriented processes playing a greater role in IAT responding compared to Associations. The primary goal of Experiment 2 was to implement a short response window IAT intended to constrain control-oriented processing, which should selectively reduce the influence of Detection and Overcoming Bias without affecting other, unrelated parameters. Though the short response window reduced Detection estimates compared to a standard race IAT, the manipulation also influenced Guessing and White-good and Black-bad Associations. The primary goal of Experiment 3 was to implement a base-rate manipulation of the proportions of good and bad attributes presented on the race IAT, which should selectively influence the Guessing parameter. The manipulation influenced the Guessing parameter in the expected direction, but it also influenced White-good and Black-bad Associations. The primary goal of Experiment 4 was to replace the race IAT face stimuli with images of reviled White people and admired Black people, which should selectively influence the Associations parameters. The counterstereotypic exemplars reduced

White-good Associations compared to the standard IAT procedure, but they also influenced Detection estimates. The primary goal of Experiment 5 was to use a known-groups paradigm with 100 White and 100 Black participants who completed a race IAT, which should selectively influence Associations. Aligning with predictions, White participants' White-good Associations were higher than Black participants' White-good Associations. However, Black-bad Associations were not different between Black and White participants. The primary goal of Experiment 6 was similar to Experiment 5, but we relied on a known groups paradigm consisting of 100 straight and 100 gay participants who responded to a sexuality IAT, which should selectively influence Associations. As expected, straight participants' Straight-good Associations were higher than gay participants' Straight-good Associations. In addition, gay participants' gay-bad Associations were lower than straight participants' gay-bad Associations. However, Detection parameter estimates were also higher among straight participants compared to gay participants.

Experiments 1–6 provide moderate evidence for the construct validity of the quad model. The quad model provided an acceptable fit to the data, and the selective-influence manipulations generally affected their intended parameters. However, many of the selective-influence manipulations also affected unintended parameters and, thus, did not demonstrate discriminant validity.

MODEL SELECTION ANALYSIS

The selective influence experiments provide evidence for the convergent validity but not the discriminant validity of the E-quad and quad models. Because both models demonstrated adequate fit to the data, we conducted a series of model-selection analyses to determine which model provided the better fit for each experiment. For each model in each experiment, we computed two Bayesian-oriented model selection indices: the deviance information criterion (DIC: Spiegelhalter et al., 2002) and the widely applicable information criterion (WAIC: Watanabe, 2010). Both of these model selection indices quantify model fit while penalizing for model complexity. As with the better-known model selection indices AIC and BIC, the model with the lowest DIC or WAIC value is preferred. In addition to index model fit, we computed w , an effect size that quantifies the magnitude of (mis)fit while accounting for the number of observations. Based on Cohen's (1992) effect size indices, we interpret a $w < 0.1$ as indicating acceptable model fit.

DIC and WAIC values are only comparable when calculated for models applied to the same data. In addition to the different Associations specification, the E-quad model also differs from the quad model by incorporating 20 practice trials from the second block of the IAT to improve model identifiability. Consequently, we employed two strategies to position ourselves to make strong comparisons between the E-quad and quad models.

The first strategy was to specify a version of the quad model that also includes the 20 practice trials from the second block of the IAT, hereafter referred to as the "original quad model." This specification corresponds to the quad model as first

proposed by Conrey et al. (2005), so it directly tests the competing assumptions of the E-quad versus quad models (i.e., good Associations for each target group versus one good and one bad Association, respectively). However, the original quad model is rarely used anymore; instead, the most commonly specified version omits practice trials (Figure 1). Consequently, this comparison has low ecological validity. Nevertheless, by comparing the E-quad and original quad model in this way, we can apply the DIC and WAIC model comparison statistics, which are widely used and easily interpreted.

Our second strategy was to specify the quad model as it is commonly used today, estimating parameters from the 120 combined critical trials and excluding practice trials. As outlined above, the E-quad model includes data from practice trials, so we cannot compare the two models using DIC or WAIC, as they are not based on the same data. Consequently, we compared the two models using the w -statistic. Chi-squared and thus T_1 , quantifies the degree to which a model (mis-) fits data, but is generally not used for model selection because it does not account for complexity in the model form. Nevertheless, in our case, the E-quad model and quad model commonly used today consist of the same number of parameters, so they are comparable in that dimension. Chi-squared also depends on the number of observations, and the E-quad model is specified based on 20 more trials per participant than the quad model (i.e., the 20 attribute practice trials). The w -statistic scales model fit (i.e., T_1) by the number of observations, and in doing so will provide a common metric by which to compare fit between the E-quad model and the quad model.

Results. Across all indices and experiments, the quad and original quad models provided a better fit to the data than did the E-quad model (Table 9). Both DIC values and WAIC values were lower for the original quad model compared to the E-quad model. We acknowledge that there is little precedent for using the w -statistic to perform model comparisons. That said, the w -statistic is lower for the quad model compared to the E-quad model in almost all experiments, which ultimately does not provide compelling evidence for the superior performance of the E-quad model.

GENERAL DISCUSSION

Guided by considerable theoretical and empirical evidence that intergroup bias primarily reflects the contributions of positive evaluations more so than negative evaluations (Brewer, 1999; Calanchini et al., 2022; Greenwald & Pettigrew, 2014; Halevy et al., 2008; Mummendey & Otten, 1998), the present research aimed to specify a variant of the quad model (Conrey et al., 2005) that estimated the contributions of only positive evaluations for each target group in an IAT. Across all experiments, the E-quad model provided an acceptable fit to both new and archival data but did not consistently demonstrate selective influence. Moreover, the quad model consistently provided a better fit to these data than did the E-quad model.

TABLE 9. Model Fit and Comparison Indices for the Quad Model and E-quad Model

Experiment	DIC		WAIC		w	
	Original Quad Model	E-quad Model	Original Quad Model	E-quad Model	Quad Model	E-quad Model
1: Temporal Stability	4474	4622	4723	4886	0.005	0.009
2: Response Deadline	4692	5060	4993	5368	0.005	0.009
3: Base Rate	4082	4354	4385	4645	0.004	0.006
4: Counterstereotypic Exemplars	3906	4082	4188	4371	0.004	0.006
5a: Known Groups (Race) <i>n</i> = 200	4302	4438	5256	5495	0.003	0.005
6a: Known Groups (Sexuality) <i>n</i> = 200	4918	5135	4613	4775	0.004	0.005
5b: Known Groups (Race) <i>n</i> = 20,000	233300	239175	233733	240965	0.0005	0.0005
6b: Known Groups (Sexuality) <i>n</i> = 20,000	265982	275298	267779	279465	0.0006	0.0005

Note. DIC: deviance information criterion; WAIC: widely applicable information criterion.

An MPT model reflects theoretical assumptions formally articulated in equation form. The response patterns associated with each processing branch reflect the researcher’s theoretical assumptions about the nature and interplay among model parameters. This logic is then formalized mathematically in a system of equations, which is used to compute model parameter estimates. As a result, an MPT model cannot demonstrate a good fit if the underlying theoretical assumptions are not supported by the observed data. Furthermore, the qualitative nature of each parameter must be demonstrated through selective-influence experiments and known-groups studies. This formal process of proposing and validating MPT models is a rigorous way to develop and refine theories.

In this manuscript, the E-quad model reflects the theoretical assumptions that intergroup bias, as operationalized in IAT responses, reflects positive evaluations of both target groups and that negative evaluations do not significantly contribute to responses. In contrast, the quad model reflects the assumptions that IAT responses reflect positive evaluations of the higher-status target group and negative evaluations of the lower-status target group, and that both groups cannot be evaluated positively. The present research provides relatively clear evidence in favor of the latter theory over the former: the model variant that accounts for positive evaluations of the higher-status target group and negative evaluations of the lower-status target group (i.e., the quad model) provided the best fit to the observed data. In turn, these findings help to advance theory about the primacy of positive versus negative evaluations in implicit social cognition (Calanchini et al., 2022; Greenwald & Pettigrew, 2014).

LIMITATIONS AND FUTURE DIRECTIONS

The present research is limited in several ways. The inconsistent selective-influence findings suggest that our manipulations may not have successfully facilitated and constrained the intended processes. For example, we replaced the White and Black faces in the standard race IAT with faces of reviled White people and admired Black people to selectively influence Associations. Though this is a classic manipulation rooted in past findings that exposure to counterstereotypic exemplars shifted implicit pro-White/anti-Black racial evaluations (Dasgupta & Greenwald, 2001), more recent research using multiple close replication studies has failed to replicate this finding. Instead, Kurdi et al. (2024) identified specific boundary conditions under which exposure to counterstereotypic exemplars might successfully influence racial evaluations. Additionally, this replication attempt suggested that the shifts in our current societal context from the historical context in which the Dasgupta and Greenwald (2001) experiments were originally conducted may be a contributing factor to the failure to replicate. Furthermore, we may reasonably suspect that some other manipulations used in this manuscript may also be similarly affected by the modern-day contexts in which we collected data. For example, all the experimental and archival data in this paper were collected online rather than in a laboratory setting, a necessity for Experiments 1–4 due to the COVID-19 pandemic. However, relying on online data collection raises further questions about the conditions under which each participant completed the experiments. Future selective influence experiments should consider how each manipulation operates within the specific experimental context.

The present research is also limited because we relied on only one paradigm: MPT modeling. Other paradigms and methods are also well-positioned to test our theory about the contributions of positive versus negative evaluations to implicit intergroup bias. For example, a researcher might reasonably administer two single-category IATs (Karpinski & Steinman, 2006), with one configured to assess White evaluations and another configured to assess Black evaluations and compare relative levels of bias towards each group. Alternatively, a researcher might reasonably administer an affect misattribution procedure (AMP; Payne et al., 2005) and separately analyze the frequencies of “more pleasant” and “less pleasant” responses to Black versus White primes. Both of these approaches could provide insight into the theory that motivated the present research. However, at the same time, both approaches are limited because summary statistics and response frequencies reflect the joint contributions of all processes that produce responses on measures like the IAT and AMP. Consequently, neither approach would offer the same depth of process-level insight into the contributions of evaluations to implicit intergroup bias as is provided by MPT modeling. Nevertheless, theories are strengthened to the extent that they are tested across multiple operationalizations, so future research investigating the contributions of positive and negative evaluations to implicit intergroup bias should explore other methodological and statistical approaches beyond what we used here.

CONCLUSION

The present research proposed the E-quad model, an alternative specification of the quad model that tests the assumption that implicit bias, as measured by the IAT, reflects the contributions of positive evaluations more than those of negative evaluations. Across a series of 6 experiments, we found that the E-quad model consistently provided an adequate fit to IAT data but responded inconsistently to theoretically guided selective influence manipulations. Moreover, model selection indices suggest that the original quad model provides a better fit to IAT data than does the E-quad model. Taken together, these experiments ultimately support the quad model as the better fit for the IAT and, in turn, support theories of intergroup bias that posit contributions from both positive and negative evaluations.

Funding statement. The authors received no specific funding for this work.

Conflict of interest statement. The authors declare no conflicts of interest.

Data availability. Preregistrations, code, data, and supplemental materials are available at <https://osf.io/5y7x3/overview>

REFERENCES

- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6(1), 57–86. <https://doi.org/10.3758/BF03210812>
- Beer, J. S., Stallen, M., Lombardo, M. V., Gonsalkorale, K., Cunningham, W. A., & Sherman, J. W. (2008). The Quadruple Process model approach to examining the neural underpinnings of prejudice. *NeuroImage*, 43(4), 775–783. <https://doi.org/10.1016/j.neuroimage.2008.08.033>
- Blair, I. V., Judd, C. M., Havranek, E. P., & Steiner, J. F. (2015). Using community data to test the discriminant validity of ethnic/racial group IATs. *Zeitschrift Für Psychologie / Journal of Psychology*, 218(1), 36–43. <https://doi.org/10.1027/0044-3409/a000006>
- Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, 86(2), 307–324. <https://doi.org/10.1037/0033-2909.86.2.307>
- Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues*, 55(3), 429–444. <https://doi.org/10.1111/0022-4537.00126>
- Calanchini, J., Lai, C. K., & Klauer, K. C. (2021). Reducing implicit racial preferences: III. A process-level examination of changes in implicit preferences. *Journal of Personality and Social Psychology*, 121(4), 796–818. <https://doi.org/10.1037/pspi0000339>
- Calanchini, J., Schmidt, K., Sherman, J. W., & Klein, S. A. W. (2022). The contributions of positive outgroup and negative ingroup evaluation to implicit bias favoring outgroups. *Proceedings of the National Academy of Sciences*, 119(40), e2116924119. <https://doi.org/10.1073/pnas.2116924119>
- Clark, K. B., & Clark, M. P. (1950). Emotional factors in racial identification and preference in Negro children. *Journal of Negro Education*, 19(3), 341. <https://doi.org/10.2307/2966491>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. <https://doi.org/10.1037//0033-2909.112.1.155>

- Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, 89(4), 469–487. <https://doi.org/10.1037/0022-3514.89.4.469>
- Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, 12(2), 163–170. <https://doi.org/10.1111/1467-9280.00328>
- Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, 81(5), 800–814. <https://doi.org/10.1037/0022-3514.81.5.800>
- De Houwer, J., & Boddez, Y. (2022). Bias in implicit measures as instances of biased behavior under suboptimal conditions in the laboratory. *Psychological Inquiry*, 33(3), 173–176. <https://doi.org/10.1080/1047840X.2022.2106755>
- Dovidio, J. F., Penner, L. A., Albrecht, T. L., Norton, W. E., Gaertner, S. L., & Shelton, J. N. (2008). Disparities and distrust: The implications of psychological processes for understanding racial disparities in health and health care. *Social Science & Medicine*, 67(3), 478–486. <https://doi.org/10.1016/j.socscimed.2008.03.019>
- Elder, J., Wilson, L., & Calanchini, J. (2023). Estimating the reliability and stability of cognitive processes contributing to responses on the Implicit Association Test. *Personality and Social Psychology Bulletin*, 50(10), 1451–1470. <https://doi.org/10.1177/01461672231171256>
- Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The Mode Model as an integrative framework. *Advances in Experimental Social Psychology*, 23, 75–109. [https://doi.org/10.1016/S0065-2601\(08\)60318-4](https://doi.org/10.1016/S0065-2601(08)60318-4)
- 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(6), 1013–1027. <https://doi.org/10.1037//0022-3514.69.6.1013>
- Gonsalkorale, K., Allen, T. J., Sherman, J. W., & Klauer, K. C. (2010). Mechanisms of group membership and exemplar exposure effects on implicit attitudes. *Social Psychology*, 41(3), 158–168. <https://doi.org/10.1027/1864-9335/a000023>
- Gonsalkorale, K., Sherman, J. W., & Klauer, K. C. (2009). Aging and prejudice: Diminished regulation of automatic race bias among older adults. *Journal of Experimental Social Psychology*, 45(2), 410–414. <https://doi.org/10.1016/j.jesp.2008.11.004>
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74(6), 1464–1480. <https://doi.org/10.1037/0022-3514.74.6.1464>
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2), 197–216. <https://doi.org/10.1037/0022-3514.85.2.197>
- Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist*, 69(7), 669–684. <https://doi.org/10.1037/a0036056>
- Halevy, N., Bornstein, G., & Sagiv, L. (2008). “In-group love” and “out-group hate” as motives for individual participation in intergroup conflict: A new game paradigm. *Psychological Science*, 19(4), 405–411. <https://doi.org/10.1111/j.1467-9280.2008.02100.x>
- Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, 50(1), 264–284. <https://doi.org/10.3758/s13428-017-0869-7>
- Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology*, 27(1), 116–159. <https://doi.org/10.1080/10463283.2016.1212966>
- Jones, K. P., Arena, D. F., Nittrouer, C. L., Alonso, N. M., & Lindsey, A. P. (2017). Subtle discrimination in the workplace: A vicious cycle. *Industrial and Organizational Psychology*, 10(1), 51–76. <https://doi.org/10.1017/iop.2016.91>
- Jones, S. C. T., Anderson, R. E., Gaskin-Watson, A. L., Sawyer, B. A., Applewhite, K.,

- & Metzger, I. W. (2020). From “crib to coffin”: Navigating coping from racism-related stress throughout the lifespan of Black Americans. *American Journal of Orthopsychiatry*, 90(2), 267–282. <https://doi.org/10.1037/ort0000430>
- Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology*, 33(1), 1–27. <https://doi.org/10.1111/j.2044-8309.1994.tb01008.x>
- Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, 25(6), 881–919. <https://doi.org/10.1111/j.1467-9221.2004.00402.x>
- Karpinski, A., & Steinman, R. B. (2006). The single category implicit association test as a measure of implicit social cognition. *Journal of Personality and Social Psychology*, 91(1), 16–32. <https://doi.org/10.1037/0022-3514.91.1.16>
- Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, 75(1), 70–98. <https://doi.org/10.1007/s11336-009-9141-0>
- Kurdi, B., Sanchez, A., Dasgupta, N., & Banaji, M. R. (2024). (When) do counterattitudinal exemplars shift implicit racial evaluations? Replications and extensions of Dasgupta and Greenwald (2001). *Journal of Personality and Social Psychology*, 126(4), 543–565. <https://doi.org/10.1037/pspa0000370>
- Lai, C. K., & Banaji, M. R. (2020). The psychology of implicit intergroup bias and the prospect of change. In D. Allen & R. Somanathan (Eds.), *Difference without domination: Pursuing justice in diverse democracies* (pp. 115–146). University of Chicago Press
- Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., . . . Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, 143(4), 1765–1785. <https://doi.org/10.1037/a0036260>
- Levy, D. J., Heissel, J. A., Richeson, J. A., & Adam, E. K. (2016). Psychological and biological responses to race-based social stress as pathways to disparities in educational outcomes. *American Psychologist*, 71(6), 455–473. <https://doi.org/10.1037/a0040322>
- Mann, T. C., Cone, J., Heggeseth, B., & Ferguson, M. J. (2019). Updating implicit impressions: New evidence on intentionality and the affect misattribution procedure. *Journal of Personality and Social Psychology*, 116(3), 349–374. <https://doi.org/10.1037/pspa0000146>
- Mathur, V. A., Harada, T., Lipke, T., & Chiao, J. Y. (2010). Neural basis of extraordinary empathy and altruistic motivation. *NeuroImage*, 51(4), 1468–1475. <https://doi.org/10.1016/j.neuroimage.2010.03.025>
- Melnikoff, D. E., & Kurdi, B. (2022). What implicit measures of bias can do. *Psychological Inquiry*, 33(3), 185–192. <https://doi.org/10.1080/1047840X.2022.2106759>
- Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, 132(2), 297–326. <https://doi.org/10.1037/0033-2909.132.2.297>
- Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, 42(1), 42–54. <https://doi.org/10.3758/BRM.42.1.42>
- Mummendey, A., & Otten, S. (1998). Positive-negative asymmetry in social discrimination. *European Review of Social Psychology*, 9(1), 107–143. <https://doi.org/10.1080/14792779843000063>
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics: Theory, Research, and Practice*, 6(1), 101–115. <https://doi.org/10.1037/1089-2699.6.1.101>
- Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18(1), 36–88.

- <https://doi.org/10.1080/10463280701489053>
- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, 105(2), 171–192. <https://doi.org/10.1037/a0032734>
- 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(3), 277–293. <https://doi.org/10.1037/0022-3514.89.3.277>
- Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The Meta-Cognitive Model (MCM) of attitudes: Implications for attitude measurement, change, and strength. *Social Cognition*, 25(5), 657–686. <https://doi.org/10.1521/soco.2007.25.5.657>
- Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., & Goel, S. (2020). A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, 4(7), 736–745. <https://doi.org/10.1038/s41562-020-0858-1>
- Pratto, F., Sidanius, J., & Levin, S. (2006). Social dominance theory and the dynamics of intergroup relations: Taking stock and looking forward. *European Review of Social Psychology*, 17, 271–320. <https://doi.org/10.1080/10463280601055772>
- Riddle, T., & Sinclair, S. (2019). Racial disparities in school-based disciplinary actions are associated with county-level rates of racial bias. *Proceedings of the National Academy of Sciences*, 116(17), 8255–8260. <https://doi.org/10.1073/pnas.1808307116>
- Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, 95(3), 318–339. <https://doi.org/10.1037/0033-295X.95.3.318>
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639. <https://doi.org/10.1111/1467-9868.00353>
- Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1
- Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M. B., Nadal, K. L., & Esquilin, M. (2007). Racial micro-aggressions in everyday life: Implications for clinical practice. *American Psychologist*, 62(4), 271–286. <https://doi.org/10.1037/0003-066X.62.4.271>
- Tajfel, H. (1982). Experimental studies of intergroup behaviour. In J.-P. Codol & J.-P. Leyens (Eds.), *Cognitive analysis of social behavior* (pp. 227–246). Springer Netherlands. https://doi.org/10.1007/978-94-009-7612-2_8
- Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, 1(2), 149–178. <https://doi.org/10.1002/ejsp.2420010202>
- Taylor, P., Kochhar, R., Fry, R., & Velasco, G. (2011). *Wealth gaps rise to record highs between Whites, Blacks and Hispanics*. Pew Research Center. <https://www.pewresearch.org/social-trends/2011/07/26/wealth-gaps-rise-to-record-highs-between-whites-blacks-hispanics/>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12), 3571–3594.