Completing a Race IAT changes racial bias

Ian Hussey & Jan De Houwer

Ghent University, Belgium

Word limit = 750 (intro & discussion), 2500 (methods & results)

Word count = 755 (intro & discussion), 1804 (methods & results)

Author Note

Corresponding author: Ian Hussey, Department of Experimental-Clinical and Health Psychology, Ghent University, Henri Dunantlaan 2, Gent 9000, Belgium. Email: ian.hussey@ugent.be. Funding provided by Ghent University grants 01P05517 to IH and BOF16/MET\_V/002 to JDH.

Abstract

The Race Implicit Association Test has been used in online studies to assess implicit racial biases in over four million participants. Results from three pre-registered experiments (*N* = 685) demonstrated that completing a Race IAT serves to change the racial attitudes that it seeks to assess. Increases in negative implicit racial attitudes were observed on two different implicit measures, but did not generalize to a behavioural measure of racial bias. Increases in positive explicit racial attitudes were also observed when meta-analyzed across experiments. Results highlight an important caveat for many forms of psychological assessment: that by measuring, we often perturb the system that we wish to understand.

Keywords: Implicit racial bias; implicit social cognition; implicit association test

Completing a Race IAT changes racial bias

All psychological testing to some degree also provides individuals with new experiences that influences their subsequent behaviour: just because a task is intended as a testing context does not mean that it will not induce behaviour. This is akin to Heisenberg’s observer effect in physics, where the act of measuring perturbs the system that we are interested in understanding. For example, there is significant literature on the question of whether asking individuals about suicidality increases the risk of future suicidal behaviour (De Cou & Schumann, 2017). In contrast, the study of implicit biases has paid relatively little attention to the question of whether assessing implicit attitudes also serves to change those attitudes. This question gains increased importance given the sheer scale of use of implicit measures: more than forty million participants Implicit Attitudes Tests (IAT) have been completed in online studies at Project Implicit ([implicit.harvard.edu](https://implicit.harvard.edu/); Xu, Nosek, & Greenwald, 2014).

Recent research has demonstrated that implicit racial attitudes can be changed in several ways, such as via evaluative conditioning and other learning paradigms (Lai et al., 2014). However, no work has examined changes in implicit attitudes due to the act of measurement. Whereas it has been examined whether completing a Race IAT influences more distal behaviours such as interpersonal warmth (Vorauer, 2012), no work to date has examined whether merely completing an IAT serves to change the very attitudes it seeks to assess. Recent evidence shows that mere completion of an IAT can establish implicit attitudes towards novel stimuli, and specifies that it does this through a known learning pathway: analogical learning due to the relational structure among the IAT’s four concept categories (Hussey & De Houwer, 2018). More specifically, when having to categorize positive and negative stimuli together with pictures of Black and White faces, participants might align the two dimensions (e.g., positive is to negative as White is to Black) thus leading to changes in liking (e.g., Black becomes more negative). We therefore examined whether completing a Race IAT (as compared to a control Flowers-Insects IAT) increased negative implicit biases against the racial out-group. Implicit bias was assessed using a Single-Category IAT (Experiment 1) and Affect Misattribution Procedure (Experiment 2). Experiment 3 also looked at changes in the shooter bias (Correll et al., 2007) which is considered to be an instance of racial behavior.

# Experiment 1

## Method

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study, and all studies in this article (either in-text or in supplementary materials: Simmons, Nelson, & Simonsohn, 2012). All inclusion and exclusion criteria, data collection stopping rules, analytic strategies and code for their implementation were pre-registered. All data and code are available in supplementary materials the OSF (<https://osf.io/7pbjq/?view_only=95883ad70c624ee780351e8aaf8044cc>), along with detailed descriptions of each measure and the full results of all models.

Sample. Participants for all studies were recruited online ([www.prolific.ac](http://www.prolific.ac)). Participants provided informed consent prior to participation in all experiments. In order to form homogenous racial in- and out-groups between the participants and stimuli we recruited white participants only. Sample size was selected based on availability of resources. No power analysis was conducted given the complexity of estimating power for interactions within mixed-effects models. 148 participants met inclusion and exclusion criteria (*M*age = 32.1, *SD* = 11.1; 47 women, 98 men, 3 identified using a non-binary category or provided no response).

Procedure and measures. Participants were randomly assigned to the Race IAT condition or the Flowers-Insects IAT condition. Participants completed the Modern Racism Scale (McConahay, 1986), then either the Race IAT or Flowers-Insects IAT, then the black faces SC-IAT and Likert ratings scale for the images of Black faces used in the IAT (1 - very negative to 7 - very positive).

Implicit Association Tests. All task parameters followed the recommendations of a methodological review the IAT (Nosek, Greenwald, & Banaji, 2005). Two versions of the IAT were employed that differed in their target stimuli. The Race IAT used the same stimuli that have been employed in the task hosted on the well-known Project Implicit website since 2002 (Xu et al., 2014). This included the target categories “Black people” and “White people” and the attribute categories “good” (positive words) and “bad” (negative words). The Flowers-Insects IAT was identical other than changing the target categories to “Flowers” (images of flowers) and “Insects” (images of insects; Greenwald, McGhee, & Schwartz, 1998).

Single-Category Implicit Association Test. A variant of the IAT, the SC-IAT contains only one target category so as to provide a procedurally non-relative measure of bias towards one category (Black people) without a contrast category (e.g., white people; Karpinski & Steinman, 2006). The presence of any learning effects between the groups could therefore be attributed to the IAT condition.

## Results

In order to increase power we employed linear mixed-effect modeling of reaction times on the SC-IAT rather than the more common strategy of using *D* scoring algorithm (Greenwald, Nosek, & Banaji, 2003). For the sake of clarity, only results of main or interaction effects testing our pre-registered hypotheses will be reported for each experiment. The full results of each model are available in the supplementary online materials.

Reaction time was entered as the dependent variable, SC-IAT block, IAT condition and their interaction were entered as fixed effects, racism was entered as a fixed-effect covariate, and participant was entered as a random effect (Wilkinson notation: RT ∼ block \* condition + racism + (1 | participant)). Results demonstrated that SC-IAT effects differed significantly between IAT conditions, *B* = 4.459, 95% CI = [1.030, 7.887], β = 0.017, 95% CI = [0.004, 0.029], *p* = .011. Inspection of the estimated marginal means revealed that participants who completed the Race IAT demonstrated more negative implicit bias towards black people on the subsequent SC-IAT than did participants who completed the Flowers-Insects IAT, as hypothesized. For illustrative purposes, differences in SC-IAT effects between groups can be quantified using traditional *D*1 scoring. This corresponds to a small difference of between the two groups, *M*diff = 0.03, 95% CI = [-0.02, 0.08], Cohen’s *d* = 0.18, 95% CI = [-0.15, 0.51].

Self-report ratings data were analyzed using a similar model, with rating as the dependent variable, IAT condition as a fixed effect, racism as a fixed-effect covariate, and participant as a random effect (rating ∼ condition + racism + (1 | participant)). No evidence was found for differences in ratings between the IAT conditions, *B* = -0.04, 95% CI = [-0.18, 0.10], β = -0.04, 95% CI = [-0.17, 0.09], *p* = .560.

# Experiment 2

In order to examine the robustness of the effect observed in Experiment 1, we conducted a close replication using a different implicit measure as a dependent variable: the Affective Misattribution Procedure (AMP: Payne, Cheng, Govorun, & Stewart, 2005). This served to decrease the procedural similarities the training task (IAT) and testing task, and therefore remove the possibility that the effects in Experiment 1 were merely a carryover effect from the training task. The design was otherwise identical to the previous experiment.

## Method

Sample. Sample size was selected by increasing relative to Experiment 1 to allow for a potentially higher attrition rate in the AMP. 213 participants met inclusion and exclusion criteria (*M*age = 35.8, *SD* = 12.1; 103 women, 108 men, 2 identified using a non-binary category or provided no response).

Procedure and measures. These were identical to Experiment 1 other than the use of the AMP instead of the SC-IAT. A single-category version of the AMP was employed so as to provide a measure of implicit racial bias towards black people in the absence of a racial contrast category (e.g., white people). Two forms of prime were used: images of black people (black primes) and grey squares (neural primes: see Payne et al., 2010). All other details of the AMP followed typical practices for this widely used implicit measure.

## Results

A binary logistic mixed-effects model was constructed with AMP ratings as the dependent variable, AMP prime type (black faces vs. neutral grey square), IAT condition and their interaction as fixed effects, racism as a fixed-effect covariate, and participant as a random effect (rating ∼ prime \* condition + racism + (1 | participant)). As hypothesized, AMP effects differed between the two IAT conditions, OR = 0.92, 95% CI = [0.90, 0.95], *p* < .001. Inspection of the estimated marginal means indicated participants who completed the Race IAT demonstrated more negative implicit bias towards images of black people on the subsequent AMP than did participants who completed the Flowers-Insects IAT. The self-report ratings data were analyzed using an identical model to Experiment 1. Results revealed no evidence for differences in self-report ratings between IAT conditions, *B* = 0.10, 95% CI = [-0.02, 0.22], β = 0.10, 95% CI = [-0.01, 0.21], *p* = .089.

# Experiment 3

In a third experiment, we examined the generalizability of this learning effect to responses in the Shooter Bias task (aka Police Officer’s Dilemma Task: Correll et al., 2007). This task presents participants with images of men who are either armed or unarmed and requires them to make “shoot” or “don’t shoot” responses under time pressure. Previous studies have demonstrated a greater propensity to shoot images of black men relative to white, therefore providing a measure of racial bias. We employed a single category version of the task that included only images of Black men. The design was otherwise identical to the previous experiments.

## Method

Participants. Sample size was selected by increasing the sample relative to the previous experiments. 246 participants met inclusion and exclusion criteria (*M*age = 36.1, *SD* = 11.5; 152 women, 91 men, 3 identified using a non-binary category or provided no response).

Procedure and measures. These were identical to Experiment 1 and 2, with the exception of the use of a Shooter Bias task as a dependent variable. All parameters of the Shooter bias task followed the recommendations of a recent methodological review (Correll, Hudson, Guillermo, & Ma, 2014). A single-category version of the Shooter Bias task was used to provide a measure of bias towards black people in the absence of a contrast category (e.g., White people) between the two IAT conditions.

## Results

We selected the three most common metrics of behaviour within the Shooter Bias task on the basis of a recent meta-analysis: differential reaction times between trial types (armed vs. not armed), response sensitivity, and response bias (Correll et al., 2014). Each was analyzed using a linear mixed effects model. First, reaction time was entered as the dependent variable, trial type (armed vs. unarmed), IAT condition and their interaction were entered as fixed effects, racism as a fixed-effect covariate, and participant as a random effect (RT ∼ trial\_type \* condition + racism + (1 | participant)). No evidence of an interaction between trial type and IAT condition was found, *B* = 0.983, 95% CI = [-0.298, 2.264], β = 0.010, 95% CI = [-0.003, 0.023], *p* = .133.

Second, we calculated an index of sensitivity, or the ability to accurately discriminate armed from unarmed individuals (d'). Differences in sensitivity between the IAT conditions were assessed in a second model: sensitivity was entered as the dependent variable, trial type (armed vs. unarmed), IAT condition as a fixed effect, racism as a fixed-effect covariate, and participant as a random effect (d' ∼ condition + racism + (1 | participant)). No evidence of differences between the two IAT conditions was found, *B* = 0.02, 95% CI = [-0.10, 0.14], β = 0.02, 95% CI = [-0.10, 0.15], *p* = .741.

Third, we calculated an index of response bias (c). This refers to participants’ biases towards proving a “shoot” response relative to a “don’t shoot” response, regardless of whether the image presented an individual who was armed or unarmed. Response bias was entered as the dependent variable, trial type (armed vs. unarmed) and experimental condition as a fixed effect, racism as a fixed-effect covariate, and participant as a random effect (c ∼ condition + racism + (1 | participant)). No evidence of differences between the two IAT conditions was found, *B* = 0.004, 95% CI = [-0.034, 0.042], β = 0.013, 95% CI = [-0.113, 0.139], *p* = .840.

The self-report ratings data were analyzed using the same model to previous experiments. In this case, differences in self-report ratings were found between the IAT condition, *B* = 0.19, 95% CI = [0.08, 0.31], β = 0.18, 95% CI = [0.08, 0.28], *p* < .001. Inspection of marginal estimated means indicated that participants rated the images of black men and women more positively when they previously completed a Race IAT than a Flowers-Insects IAT. Finally, given their similarity across the three experiments, all self-report ratings were submitted to a random effects meta-analysis. This model was identical to the previous analyses of self-report ratings other than also including experiment as a random effect, but was not pre-registered. Results were consistent with those from Experiment 3: participants ratings of images of black people’s faces were more positive when they previously completed a Race IAT than a Flowers-Insects IAT, *N* = 603, *B* = 0.10, 95% CI = [0.03, 0.17], β = 0.10, 95% CI = [0.03, 0.16], *p* = .004.

# Discussion

Results from three pre-registered studies demonstrated that merely completing a Race IAT served to change the attitudes it intends to assess. Experiments 1 and 2 found that completing a Race IAT (compared to a flower-insects IAT) increased negative implicit negative racial bias towards black people on a subsequent implicit measure (the SC-IAT and AMP). However, Experiment 3 found no evidence for the generalizability of this effect to the Shooter Bias task.

The direction of the effects on the implicit measures (Experiments 1 and 2) was consistent with previous work suggesting that the IAT can serve as an analogical learning context due to the relational structure among its four categories (pairs of opposites, e.g., white:black::positive:negative: Hussey & De Houwer, 2018). Unexpectedly, however, results from Experiment 3 and a meta-analysis of all three experiments demonstrated the opposite pattern of effect on the self-report measures: completing a Race IAT increased positive explicit evaluations of black people relative to completing a non-racial Flowers-Insects IAT. These opposing results may be due to a decreased willingness to self-disclose racial biases (Vorauer, 2012).

Future research should examine whether these learning effects due to competing the Race IAT persist across time. Recent evidence demonstrates that even interventions intended to change implicit attitudes do not show sustained changes over time (Lai et al., 2014, 2016). It may be the case that the learning effects observed here are also temporary in nature. It would also be interesting to look at individual differences in the susceptibility to these learning effects. Future research might also examine whether the relative status of racial in- and out-group pairs influences Race IAT learning effects. In the current experiments, all three experiments employed homogenous in- and out-groups by recruiting white participants assessing evaluations black people. Previous research has demonstrated that implicit in-group bias is related to a group’s relative social status (cf. Rudman, Feinberg, & Fairchild, 2002). It is therefore plausible learning effects due to the Race IAT are also influenced by the relative status of racial in- and out-groups. Any such differences in the malleability of racial biases could help direct future effects to change such biases, either by providing a new metric of particularly rigid implicit attitudes or by highlighting particularly amenable targets for intervention.

References

Correll, J., Hudson, S. M., Guillermo, S., & Ma, D. S. (2014). The Police Officer’s Dilemma: A Decade of Research on Racial Bias in the Decision to Shoot. *Social and Personality Psychology Compass*, *8*(5), 201–213. https://doi.org/10.1111/spc3.12099

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, *92*(6), 1006–1023. https://doi.org/10.1037/0022-3514.92.6.1006

De Cou, C. R., & Schumann, M. E. (2017). On the Iatrogenic Risk of Assessing Suicidality: A Meta-Analysis. *Suicide and Life-Threatening Behavior*. https://doi.org/10.1111/sltb.12368

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (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

Hussey, I., & De Houwer, J. (2018). The Implicit Association Test as an analogical learning task. *Experimental Psychology*, Advance online publication.

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

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., … 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

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016. https://doi.org/10.1037/xge0000179

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, Discrimination, and Racism* (pp. 91–125). San Diego, CA: Academic Press.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality & Social Psychology Bulletin*, *31*(2), 166–180. https://doi.org/10.1177/0146167204271418

Payne, 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

Payne, K., Krosnick, J. A., Pasek, J., Lelkes, Y., Akhtar, O., & Tompson, T. (2010). Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology*, *46*(2), 367–374. https://doi.org/10.1016/j.jesp.2009.11.001

Rudman, L. A., Feinberg, J., & Fairchild, K. (2002). Minority Members’ Implicit Attitudes: Automatic Ingroup Bias As A Function Of Group Status. *Social Cognition*, *20*(4), 294–320. https://doi.org/10.1521/soco.20.4.294.19908

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21 word solution*. Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2160588

Vorauer, J. D. (2012). Completing the Implicit Association Test Reduces Positive Intergroup Interaction Behavior. *Psychological Science*, *23*(10), 1168–1175. https://doi.org/10.1177/0956797612440457

Xu, K., Nosek, B., & Greenwald, A. (2014). Psychology data from the Race Implicit Association Test on the Project Implicit Demo website. *Journal of Open Psychology Data*, *2*(1). https://doi.org/10.5334/jopd.ac