Completing a Race IAT increases implicit racial bias

Ian Hussey & Jan De Houwer

Ghent University, Belgium

Author Note

Correspondence should be addressed to Ian Hussey, Department of Experimental-Clinical and Health Psychology, Ghent University, Henri Dunantlaan 2, Gent 9000, Belgium. Email: ian.hussey@ugent.be. The preparation of this article was made possible by Ghent University grants 01P05517 to IH and BOF16/MET\_V/002 to Jan De Houwer.

Abstract

Keywords:

Completing a Race IAT increases implicit racial bias

all psychological testing to some degree also provides individuals with new experiences.

This is akin to Heisenburg’s observer effect, whereby the act of testing perturbs the system.

This is acknowledged in many clinical domains. We acknowledge this in ethical review proposals: just because a task is intended as a testing context, does not mean that it will not train or induce other experiences or behaviour. For example, the fields of suicide and trauma have examined whether asking about suicide increases the risk of future attempts (De Cou & Schumann, 2017), whether asking about trauma is distressing (Jaffe, DiLillo, Hoffman, Haikalis, & Dykstra, 2015).

The implications of these results for future research of course depend on any learning effects direction and magnitude. (e.g., lowering the risk of suicide would be desirable, raising it would be problematic).

Less attention has been paid to iatrogenic effects – whereby one induces the thing that one attempts to measure - in domains of social psychology/the implicit measures such as the IAT.

[or more broadly to implicit measures – check muhlenkamp and the cha studies]

The Race IAT is used to assess racial biases, millions tested [Xu studies].

Implicit racial attitudes can be intentionally manipulated in several ways, such as via evaluative conditioning and other ways of arranging the environment (Lai et al., 2014). No work has examined iatrogenic effects for the IAT itself. [tone down the Lai references and keep it about the measure]

Importantly,

A single block of IAT categorisations can induce learning (ebert)

the act of completing an IAT has recently been shown to be capable of establishing attitudes towards novel stimuli (Hussey & De Houwer, REF).

specifically, the race IAT has also been shown to induce.

However, it’s not clear what mechanism might drive this, and/or completing the IAT changes implicit measures themselves.

We therefore examined whether the Race IAT can increase negative implicit biases against the racial out-group.

# Experiment 1

## Method

We adopted Simmons, Nelsons and Simonsohn’s (2012) “21 Word Solution” transparency statement: 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. All inclusion and exclusion criteria, data collection stopping rules, hypotheses analytic strategies, and code for implementation were pre-registered. This includes all details of the mixed-effects models (see below). We also employed both open measures and open data (XXX).

Sample. All experiments recruited participants through the Prolfiic platform ([www.prolific.ac](http://www.prolific.ac)) and were completed online through participants’ internet browsers. Inclusion criteiria were white ethinicity (to create a homogenous in-group), age 18 to 65, English as a first language, full use of both hands, normal or corrected to normal vision, participant rating in previous experiments on Prolific ≥ 90%, and no participation in the researchers’ previous experiments. Exclusion criteria were incomplete data on any task, or more than 10% of trials on the IAT or SC-IAT with reaction times < 300 ms. Participants were paid £1.20 for their participation, and provided informed consent prior to participation. 152 individuals participated (*M*age = 32.00, *SD* = 11.03; 48 women, 101 men, 1 identified as non-binary and 2 did not provide data). After applying exclusion criteria 145 participants remained in the analytic sample.

Procedure and measures. Participants were randomly assigned to the Race (induction) condition or the Flowers-Insects (control) condition when they began the experiment. Participants completed the Modern Racism Scale, either the Race IAT or Flowers-Insects IAT, and then the black faces SC-IAT and ratings scales. IAT and SC-IAT block order was counterbalanced between participants, as was the order of the SC-IAT and ratings scales.

Modern racism scale. This seven-item scale includes items such as “Black people are getting too demanding in their push for equal rights” and uses a five point response scale (strongly disagree to strongly agree: McConahay, 1986).

Implicit Association Tests. The IAT assesses the relative speed with which participant can categorize two target categories (black people and white people) and two attribute categories (good and bad). It does so by comparing how quickly participants respond when one set of targets and attributes share a response key (e.g., press left for black people or bad, press right for white people or good) with how quickly they respond when intersections are reversed (e.g., press left for black people or good, press right for white people or bad). The 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 race IAT hosted on the well-known Project Implicit website since XX (year?) (Xu, Nosek, & Greenwald, 2014). This employed the target categories “black people” (six pictures of black men and women’s faces) and “white people” (six pictures of white men and women’s faces), and the attribute categories “good” (joy, happy, laughter, love, glorious, pleasure, peace, and wonderful) and “bad” (evil, agony, awful, nasty, terrible, horrible, failure, and hurt). The Flowers-Insects IAT was identical other than changing the target categories to “Flowers” (six pictures of flowers) and “Insects” (six pictures of insects).

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 task employed three blocks of trials (block 1: 10 trials; block 2: 70 trials; block 3: 70 trials). Blocks 2 and 3 each presented the categories an unequal number of times so as to provide a roughly equal number of left and right responses (e.g., left response: 20 black people trials & 20 good trials; right response: 30 bad trials). Only data from the SC-IAT’s critical blocks (2 and 3) were analyzed.

Ratings scale. Participants rated the images of black men and women’s faces used in the race IAT using a seven-point scale (very negative to very positive).

## Results

Analytic strategy. Although typically used as a testing task, the IATs were used as training tasks here. As such, only descriptive statistics for accuracy and latency performances on the IAT are reported here. Typically, responses on the IAT and SC-IAT would be quantified and standardized using a scoring algorithm such as the *D* score (Greenwald, Nosek, & Banaji, 2003), which has been shown to limit the influence of outliers reaction times and control for general responding speed between participants. Under the expectation that the effect size we would observe here would be quite small, we elected to employ an alternative (pre-registered) strategy to *D* scoring that would provide greater statistical power while still limiting the influence of outliers and controlling for general responding speed. Specifically, reaction times on the SC-IAT test blocks (2 & 3) were taken as the raw data. Reaction times that deviated from the mean by > 2.5 standard deviations and were removed as outliers (0.55% trials removed). Data were then entered into a linear mixed-effects model (REF). These provide greater power to detect effects by considering all data points generated by each participant (e.g., 140 reaction times within the SC-IAT’s critical blocks), while acknowledging the non-independence of the multiple reaction times produced by each participant (i.e., the hierarchical nature of the data) and therefore controlling for differences in general responding speed between participants. Reaction time was entered as the dependent variable, SC-IAT block and experimental condition were entered as fixed effects (including interactions), racism was entered as a fixed-effect covariate, and participant was entered as a random effect. For the sake of precision, the model was defined in Wilkinson notation as: RT ∼ block \* condition + racism + (1 | participant). Our hypothesis that the SC-IAT effect would differ between the two IAT conditions was tested by the interaction effect between SC-IAT block and experimental condition. For the sake of brevity, only results relating to the key, pre-registered hypotheses will be reported. Results of the full model are included in the supplementary materials on the OSF. Results demonstrated that reaction time was predicted by the interaction between IAT block and experimental condition, *B* = 4.46, 95% CI = [1.03, 7.89], β = 0.02, 95% CI = [0.00, 0.03], *p* = .011. Inspection of the estimated marginal means indicated that the effect was in the predicted direction: that is, participants who completed the Race IAT demonstrated more negative implicit bias towards images of black people on the subsequent SC-IAT than did participants who completed the (control) Flowers-Insects IAT (see Table/Figure XXX). Full results of the model, and all subsequent models, can be found in the supplementary materials.

The self-report ratings data was submitted to a similar linear mixed-effects model: rating was entered as the dependent variable, experimental condition was entered as a fixed effect, racism was entered as a fixed-effect covariate, and participant was entered as a random effect: rating ∼ condition + racism + (1 | participant). This random effect acknowledged the non-independence of the multiple ratings provided by each participant. Our hypothesis that ratings differed between the two IAT conditions referred to the main-effect for experimental condition. Results revealed no significant main-effect for condition, *B* = 0.04, 95% CI = [-0.10, 0.18], β = -0.04, 95% CI = [-0.09, 0.17], *p* = .560.

# Experiment 2

One potential limitation of Experiment 1 was the procedural overlap between the training task (IAT) and testing task (SC-IAT). It may be the case that the observed effects were a carryover effect from the training task rather than representing changes in implicit racial bias. Experiment 2 therefore employed a different implicit measure that shares little procedural overlap with the IAT: the Affective Misattribution Procedure (Payne, Cheng, Govorun, & Stewart, 2005). The design was otherwise identical to the previous experiment, including use of pre-registration.

## Method

Participants. 120 individuals provided some data. 100 participants (46 women, 54 men, *M*age = 34.34, *SD* = 12.37) provided complete data and were included in the analysis.

Procedure and measures. These were identical to Experiment 1, with the exception of the use of an AMP over a SC-IAT as the dependent variable. The AMP [insert explanation of AMP procedure here]. 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 contrast category (e.g., white people). Following previous research, two forms of prime were used: images of black people (black primes) and grey squares (neural primes: see Payne REF). Participants also provided a self-report measure of stimulus awareness after the AMP. This asked whether the images that were presented in the AMP were of a) black people, b) white people, c) both, or d) I don’t know. This measure was included in an exploratory fashion and not included in our data analysis plan, and therefore is not reported here (see online materials).

## Results

Outlier removal removed (N RTS, 2.21%).

AMP (*M*RT = 634, *SD* = 114)

Ratings of the target stimuli as either positive or negative on the AMP were submitted to a logistic mixed-effects model. Ratings were entered the dependent variable, AMP prime type (black faces vs. neutral grey square) and experimental condition were entered as fixed effects (including interactions), racism was entered as a fixed-effect covariate, and participant was entered as a random effect: rating ∼ prime \* condition + racism + (1 | participant). Our hypothesis that the AMP effect differed between the two IAT conditions referred to the interaction effect between AMP prime type and experimental condition. This interaction effect was found to be significant, OR = 0.92, 95% CI = [0.90, 0.95], *p* < .001. Inspection of the estimated marginal means indicated that the effect was in the predicted direction: 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 (control) Flowers-Insects IAT (see Table/Figure XXX).

The self-report ratings data was submitted to an identical analysis to the previous experiment. Again, results revealed no significant main-effect for condition on self-report ratings, *B* = -0.10, 95% CI = [-0.22, 0.02], β = -0.10, 95% CI = [-0.21, 0.01], *p* = .089.

Self-report ratings from Experiments 1 and 2 were submitted to a meta-analysis within an additional, exploratory, linear mixed-effects model. This model was similar to previous ones but also included experiment as a random effect. Again, results revealed no significant main-effect for condition on self-report ratings, *B* = -0.02, 95% CI = [-0.12, 0.07], β = -0.02, 95% CI = [-0.11, 0.06], *p* = .601.

# Discussion

Across two pre-registered experiments, we demonstrate that white participants complete a Race IAT demonstrate more negative implicit negative racial bias towards black people on a subsequent implicit measure (SC-IAT and AMP). No evidence that the Race IAT’s impacted explicit ratings of black people was observed.

Attempts to decrease implicit racial out-group biases have been shown not to have lasting effects (Lai et al., 2014, 2016). It is likely that the effects observed here likewise have little or no long term influence.

Limitations: the self-report task was asked people to rate the images of black people rather than about racial attitudes specifically. Future studies might employ a measure of racism as a DV after the IAT, rather than a covariate completed first as done here.

Issues raised: magnitude of the learning effect (small); persistence of effect across time (likely not); specificity of effect to the IAT (likely not); explanation of the effect (analogical, based on previous work).

References

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., 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

Jaffe, A. E., DiLillo, D., Hoffman, L., Haikalis, M., & Dykstra, R. E. (2015). Does it hurt to ask? A meta-analysis of participant reactions to trauma research. *Clinical Psychology Review*, *40*, 40–56. https://doi.org/10.1016/j.cpr.2015.05.004

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. B., 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

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

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

SC IAT

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 95% CI | |  | 95% CI | |  |  |  |
|  | B | lower | upper | std Beta | lower | upper | F | df | *p* |
| Intercept | 662.79 | 601.30 | 724.28 |  |  |  |  |  | < .001 |
| Block | -0.99 | -4.42 | 2.44 | 0.00 | -0.02 | 0.01 | XXX | XXX | .57 |
| Condition | -7.60 | -24.30 | 9.11 | -0.03 | -0.09 | 0.03 | XXX | XXX | .37 |
| Block \* condition | 4.46 | 1.03 | 7.89 | 0.02 | -0.05 | 0.08 | XXX | XXX | .01 |
| Racism | 0.72 | -2.56 | 4.00 | 0.01 | 0.00 | 0.03 | XXX | XXX | .67 |

Ratings exp 1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 95% CI | |  | 95% CI | |  |  |  |
|  | B | lower | upper | std Beta | lower | upper | F | df | *p* |
| Intercept | 662.79 | 601.30 | 724.28 |  |  |  |  |  | < .001 |
| Block | -0.99 | -4.42 | 2.44 | 0.00 | -0.02 | 0.01 | XXX | XXX | .57 |
| Condition | -7.60 | -24.30 | 9.11 | -0.03 | -0.09 | 0.03 | XXX | XXX | .37 |
| Block \* condition | 4.46 | 1.03 | 7.89 | 0.02 | -0.05 | 0.08 | XXX | XXX | .01 |
| Racism | 0.72 | -2.56 | 4.00 | 0.01 | 0.00 | 0.03 | XXX | XXX | .67 |

AMP

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | 95% CI | |  |
|  | Odds ratio | lower | upper | *p* |
| Intercept | 5.18 | 2.06 | 13.03 | < .001 |
| Prime type | 1.04 | 1.00 | 1.07 | .03 |
| Condition | 1.12 | 0.91 | 1.36 | .29 |
| Prime type \* condition | 0.92 | 0.90 | 0.95 | < .001 |
| Racism | 0.96 | 0.91 | 1.01 | .10 |

Ratings exp 2

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 95% CI | |  | 95% CI | |  |  |  |
|  | B | lower | upper | std Beta | lower | upper | F | df | *p* |
| Intercept | 662.79 | 601.30 | 724.28 |  |  |  |  |  | < .001 |
| Block | -0.99 | -4.42 | 2.44 | 0.00 | -0.02 | 0.01 | XXX | XXX | .57 |
| Condition | -7.60 | -24.30 | 9.11 | -0.03 | -0.09 | 0.03 | XXX | XXX | .37 |
| Block \* condition | 4.46 | 1.03 | 7.89 | 0.02 | -0.05 | 0.08 | XXX | XXX | .01 |
| Racism | 0.72 | -2.56 | 4.00 | 0.01 | 0.00 | 0.03 | XXX | XXX | .67 |