Completing a Race IAT increases implicit racial bias

Ian Hussey & Jan De Houwer

Ghent University, Belgium

Word limit: intro & discussion = 750, methods & results = 2500

Word count: intro & discussion = 318+427=745, methods & results = 1331

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 Implicit Association Test has been used in online studies to assess implicit racial attitudes in over seven million participants. Although typically used as an assessment measure, results from four pre-registered experiments (*N* = 940) demonstrated that completing a Race IAT exacerbates the negative implicit attitudes that it seeks to assess. Increases in White participants’ negative automatic racial evaluations of Black people were observed across two different implicit measures (SC-IAT and AMP) but did not generalize to another measure of automatic racial bias (Shooter Bias task). Results highlight an important caveat for the Race IAT, but also for many other 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 increases implicit racial bias

All psychological testing provides individuals with new experiences that influence their subsequent behavior, akin to Heisenberg’s (1958) observer effect in physics where the act of measuring perturbs the system being studied. Although there is much research on whether implicit attitudes can be purposefully changed via intervention (Lai et al., 2014), the study of racial attitudes has paid little attention to the question of whether there merely measuring implicit attitudes serves to change them (cf. Vorauer, 2012). This seems strange given how frequently and widely measures of implicit racial attitudes are used: more than seven million Implicit Association Tests (IATs) have been completed on the Project Implicit website (implicit.harvard.edu; Xu, Nosek, & Greenwald, 2014), and the subject has become sufficiently mainstream to be a topic in US presidential debates (Yudkin & Bavel, 2018). It is also in stark comparison to other domains that have examined comparable questions (e.g., whether asking about suicide increases risk: Blades, Stritzke, Page, & Brown, 2018).

Recent evidence demonstrated that completing an IAT can establish implicit evaluations toward entirely novel stimuli. This was suggested to occur due to analogical learning, which takes place based on the relational structure among the IAT’s four concept categories that form pairs of opposites (Hussey & De Houwer, 2018; see Gentner & Smith, 2013). Here, we examine whether completing an IAT can also change well-established implicit racial attitudes. Specifically, when people have to categorize positive and negative stimuli together with pictures of Black and White faces, they might align the two dimensions analogously (e.g., “White people are to Black people as positive is to negative”) so that any existing positive in-group evaluation also leads to the inference that the out-group is negative. Across four experiments we examined whether completing a Race IAT compared to a control Flowers-Insects IAT increased negative evaluations of the racial out-group on subsequent measures of automatic behaviour (Single-Category IAT, Affect Misattribution Procedure, or Shooter Bias task).

# Method

All inclusion and exclusion criteria, data collection stopping rules, analytic strategies and code for their implementation were pre-registered. Preregistrations, data and code for both measures and analyses are available on the OSF (<https://osf.io/7pbjq/?view_only=95883ad70c624ee780351e8aaf8044cc>). Detailed descriptions of each measure and the full results of all models are available in the Supplementary Materials.

## **Samples**

On the basis that the majority of Race IATs are conducted using Internet samples (see Xu et al., 2014), participants for all studies were recruited online ([www.prolific.ac](http://www.prolific.ac)) and experiments were completed in participant’s web browser. 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 only participants of White race, and dependant variables employed images of Black people. Sample sizes and were selected and pre-registered based on availability of resources and supported by the results of a Monte-Carlo simulation study for the replication study (see Supplementary Materials).

Procedure and measures. Participants completed the Modern Racism Scale (McConahay, 1986), a either a Race IAT or a Flowers-Insects IAT, then completed a second single-category measure of automatic behaviour towards Black people which varied between studies, and finally provided self-report ratings of the images of Black faces used in the behavioural tasks (1 to 7 Likert scale from very negative to very positive). The presence of any differences between the conditions could therefore be attributed to the influence of the IAT’s use of racial or non-racial stimuli.

All IAT task parameters followed the recommendations of a methodological review the IAT (Nosek, Greenwald, & Banaji, 2005). 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). The Flowers-Insects IAT was identical to this other than changing the concept categories from images of Black and White people to images of Flowers and Insects (Greenwald, McGhee, & Schwartz, 1998).

Changes in automatic behaviour due to completing the IAT were assessed with subsequent task: the Single-Category IAT (Karpinski & Steinman, 2006), Affective Misattribution Procedure (Payne, Cheng, Govorun, & Stewart, 2005), or Shooter Bias task (Correll, Hudson, Guillermo, & Ma, 2014). In all cases, we employed single-category variants of these tasks that included images of Black people (but not White people) so as to provide a procedurally non-relative measures of implicit evaluations of the racial out-group only. All other details of each measure followed typical practices for these widely used measures and were based on the result of methodological reviews where possible (Correll et al., 2014; Nosek et al., 2005). The Single-Category IAT (Experiment 1) provided measure of automatic evaluations towards Black people through the relative speed of categorization of Black faces when mapped to the same key as either positive or negative words (Karpinski & Steinman, 2006). The AMP (Experiments 2 & 4) provided a measure of automatic evaluation of Black people through the misattribution of the valence of prime stimuli (Black faces vs. neutral primes) to target stimuli (Chinese characters; see Payne et al., 2010 for previous use of the single-category AMP). Both the SC-IAT and AMP employed the same images of Black faces as the Race IAT. Finally, the Shooter Bias task (Experiment 3) captured automatic racial bias by presenting participants with images of Black men who are either armed with a gun 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 (Correll et al., 2014).

# Results

## Experiment 1

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), 11 participants were excluded. In line with our pre-registered data analysis plan, we employed linear mixed effects modeling of reaction times on the SC-IAT. This approach maximizes power by taking all 120 reaction times from each participant into account, but also accomplishes the same goals as the more common strategy of *D* scoring (Greenwald, Nosek, & Banaji, 2003). 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. As hypothesized, SC-IAT effects differed between the two IAT conditions, *B* = 4.459, 95% CI [1.030, 7.887], β = 0.017, 95% CI [0.004, 0.029], *p* = .011, with participants in the Race IAT condition demonstrating more negative implicit evaluations of the Black people than the Flowers-Insects IAT condition. No differences were found in self-report ratings of the Black faces 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

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), 19 participants were excluded. AMP effectcs were analysed using a logistic mixed-effects model with AMP ratings as the dependent variable, AMP prime type (Black faces vs. a neutral grey square), IAT condition, and their interaction as fixed effects, racism as a fixed-effect covariate, and participant as a random effect. As hypothesized, AMP effects differed between the two IAT conditions, OR = 0.92, 95% CI [0.90, 0.95], *p* < .001, with participants in the Race IAT condition demonstrating more negative implicit evaluations of the Black people than the Flowers-Insects IAT condition. Again, no difference in self-report ratings were found between the IAT conditions, *B* = 0.10, 95% CI [-0.02, 0.22], β = 0.10, 95% CI [-0.01, 0.21], *p* = .089.

## Experiment 3

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), 48 participants were excluded. Effects on the Shooter Bias task are typically analyzed using multiple metrics. We selected the three most common on the basis of a recent meta-analysis (Correll et al., 2014): differential reaction times between trial types (armed vs. not armed), response sensitivity (*d*' scores: the ability to accurately discriminate armed from unarmed individuals), and response bias (*c* scores: propensity towards “shoot” responses relative to a “don’t shoot” responses using the metric). Each was analyzed using a linear (mixed effects) model that compared scores between IAT conditions while controlling for self-reported racism. No evidence of the key effect was found on the three metrics (reaction times: interaction between trial type and IAT condition, *B* = 0.983, 95% CI [-0.298, 2.264], β = 0.010, 95% CI [-0.003, 0.023], *p* = .133; sensitivity: main effect for IAT condition, *B* = 0.02, 95% CI [-0.10, 0.14], β = 0.02, 95% CI [-0.10, 0.15], *p* = .741; response bias: main effect for IAT condition, *B* = 0.004, 95% CI [-0.034, 0.042], β = 0.013, 95% CI [-0.113, 0.139], *p* = .840). Self-report ratings of Black people’s faces were more positive when they previously completed a Race IAT than a Flowers-Insects IAT, *B* = 0.19, 95% CI = [0.08, 0.31], β = 0.18, 95% CI [0.08, 0.28], *p* < .001.

## Experiment 4

Finally, we conducted an exact self-replication of Experiment 2, using the AMP as the dependent variable, in order to assess the replicability of the effect on implicit measures of implicit evaluation. 333 participants met inclusion and exclusion criteria (*M*age = 36.2, *SD* = 12.6; 178 women, 156 men), 43 participants were excluded. As hypothesized and consistent with Experiment 2, AMP effects differed between the two IAT conditions, OR = 0.94, 95% CI [0.92, 0.97], *p* < .001, with participants in the Race IAT condition demonstrating more negative implicit evaluations of the Black people than the Flowers-Insects IAT condition. Self-report ratings of Black people’s faces were more positive when they previously completed a Race IAT than a Flowers-Insects IAT, *B* = 0.08, 95% CI = [0.08, 0.28], β = 0.17, 95% CI [0.08, 0.25], *p* = .004.

# Discussion

Results from four studies demonstrated that White participants who completed a Race IAT showed larger negative automatic evaluations of Black people on a subsequent implicit measure compared to those who completed a Flower-Insects IAT. Merely completing a Race IAT therefore changes the implicit attitudes it intends to assess. This result was observed when the outcome measure did (SC-IAT) or did not (AMP) share procedural similarities with the IAT, suggesting that the effects were not simply a carryover effect from the training task. Additionally, by controlling for differences in self-reported baseline racism, results were not due to baselines differences between groups. This learning effect did not appear to generalize to another measure of non-evaluative automatic racial behaviour (Shooter Bias task), suggesting the boundaries of the effect’s generalizability. Finally, a successful direct self-replication demonstrated the effect’s robustness. Mixed results were found across studies for change on the self-report ratings of Black individuals and were therefore not over interpreted, but may speak to the general pattern of the limitations of self-reports in socially sensitive domains due to demand characteristics (Nosek, 2007).

Results add to findings elsewhere suggesting that the IAT functions as a training as well as a testing task, which can not only establish automatic evaluations towards novel stimuli but also change existing automatic racial evaluations (Hussey & De Houwer, 2018). Furthermore, changes in automatic evaluations of the racial out-group were consistent with the idea that this was due to the analogical relational structure among the IAT’s four concept categories (e.g., “White is to Black as positive is negative”).

Evidence suggests that attempts to change implicit attitudes do not show lasting effects over time (Lai et al., 2014, 2016). The learning effect observed here is therefore also likely to be temporary in nature. However, even temporary effects have important consequences: previous research has suggested that the order in which implicit and explicit measures are administered can have an impact on their respective results (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005; Nosek et al., 2005), yet has not suggested the learning pathway via which this occurs. Our results provide an explanation: completing an IAT influenced the attitudes it wishes to assess, and may therefore influence responses on a subsequent task or the association between performances on different tasks. These results highlight a more general need to consider the magnitude and direction of training effects induced by psychological assessment measures. Future research could examine individual differences in susceptibility to these learning effects (e.g., based on analogical ability or the tendency to search for deeper structure; Cacioppo, Petty, & Kao, 1984).

References

Blades, C. A., Stritzke, W. G. K., Page, A. C., & Brown, J. D. (2018). The benefits and risks of asking research participants about suicide: A meta-analysis of the impact of exposure to suicide-related content. *Clinical Psychology Review*. https://doi.org/10.1016/j.cpr.2018.07.001

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307. https://doi.org/10.1207/s15327752jpa4803\_13

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

Gentner, D., & Smith, L. A. (2013). Analogical learning and reasoning. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology* (1st ed., pp. 668–681). New York, NY: Oxford University Press.

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

Heisenberg, W. (1958). Physics and philosophy.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A Meta-Analysis on the Correlation Between the Implicit Association Test and Explicit Self-Report Measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613

Hussey, I., & De Houwer, J. (2018). The Implicit Association Test as an analogical learning task. *Experimental Psychology*, Accepted manuscript. https://doi.org/10.31234/osf.io/yug9h

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. (2007). Implicit–Explicit Relations. *Current Directions in Psychological Science*, *16*(2), 65–69. https://doi.org/10.1111/j.1467-8721.2007.00477.x

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

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, F. K., Nosek, B. A., & Greenwald, A. G. (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

Yudkin, D. A., & Bavel, J. V. (2018, January 20). The Roots of Implicit Bias. *The New York Times*. Retrieved from https://www.nytimes.com/2016/12/09/opinion/sunday/the-roots-of-implicit-bias.html