nature human behaviour

Registered Report

https://doi.org/10.1038/s41562-026-02439-y

Challenging the mechanism for the implicit association test

Received: 21 April 2023

Accepted: 2 March 2026

Published online: 16 April 2026

Check for updates

Kyle J. LaFollette[1,2] ✉, Doroteja Rubez[2], Heath A. Demaree[2] & Amit Goldenberg[3,4,5]

Implicit biases are stereotypes and attitudes that influence decisions and actions, contributing to discrimination and societal inequities. The implicit association test is the most widely used tool for measuring implicit bias, assessing response time in sorting stimuli into labelled categories. Most interpretations assume that implicit association test performance (*D*-scores) reflects conflicting associative memories or decision ease. We challenged this assumption by decomposing *D*-scores into additional cognitive processes that may influence results, particularly response caution—the tendency to trade speed for accuracy. Using racing diffusion models across 39 topics ($N = 115{,}601$), we found that response caution explained significantly more variance in *D*-scores beyond decision ease. Response caution also best predicted explicitly reported biases. These findings challenge the traditional interpretation of *D*-scores as primarily reflecting associative memory activation and highlight the need to consider multiple cognitive processes when assessing implicit biases.

Implicit biases are attitudes or stereotypes that affect our understanding, actions and decisions, often without our knowledge[1]. Social scientists typically define implicitness as a property of unconscious mental representations or of measures that are designed to test constructs indirectly, avoiding the filter of self-report[2–5]. These features of implicit bias make its measurement very challenging. To overcome this challenge, social scientists are tasked with developing measures that capture implicit behaviour and that are less susceptible to external pressures and social desirability than self-report[1,6–9]. Developing these measures is especially difficult, because people seem to be able to fake their performance even when measures are supposed to be implicit[10–13].

Many different measures have attempted to capture implicit bias, but the most commonly used one is the implicit association test (IAT)[3,4]. The IAT involves presenting participants with target concepts (for example, images of Black or white faces) and attribute concepts (for example, good or bad words) and asking the participants to sort these stimuli into their respective categories (Fig. 1). Target and attribute categories are later combined such that they are either compatible with biased beliefs (that is, white/good and Black/bad) or incompatible (that is, white/bad and Black/good). In the incompatible block of the race IAT, as one prominent example, images of Black faces and positive or good words (for example, 'beautiful' or 'pleasant') are mapped onto one response key, while images of white faces and negative or bad words (for example, 'ugly' or 'unpleasant') are mapped onto another key. The compatible block has the opposite configuration (that is, white/good and Black/bad). People tend to respond more slowly on bias-incompatible blocks than on bias-compatible blocks, a behaviour termed the 'IAT effect'. To quantify this effect, many researchers use *D*-scores, which are the difference in response time between compatible and incompatible blocks divided by their pooled standard deviation. At least six *D*-score algorithms have been proposed, each differing in their treatment of short response times and error trials[14,15].

Given the assumed ability of the IAT to bypass external pressures and social desirability[3–5], it has been widely used to detect and evaluate bias in healthcare[16,17], law enforcement[18] and education[19,20]. Although the IAT has been widely adopted, it has also been heavily criticized. One critique of the IAT is its limited construct validity and general inability to predict real-world behaviours[5,21,22]. Different

[1]Booth School of Business, University of Chicago, Chicago, IL, USA. [2]Department of Psychological Sciences, Case Western Reserve University, Cleveland, OH, USA. [3]Harvard Business School, Harvard University, Boston, MA, USA. [4]Department of Psychology, Harvard University, Cambridge, MA, USA. [5]Digital, Data and Design Institute, Harvard University, Boston, MA, USA. ✉e-mail: kyle.lafollette@chicagobooth.edu

https://doi.org/10.1038/s41562-026-02439-y

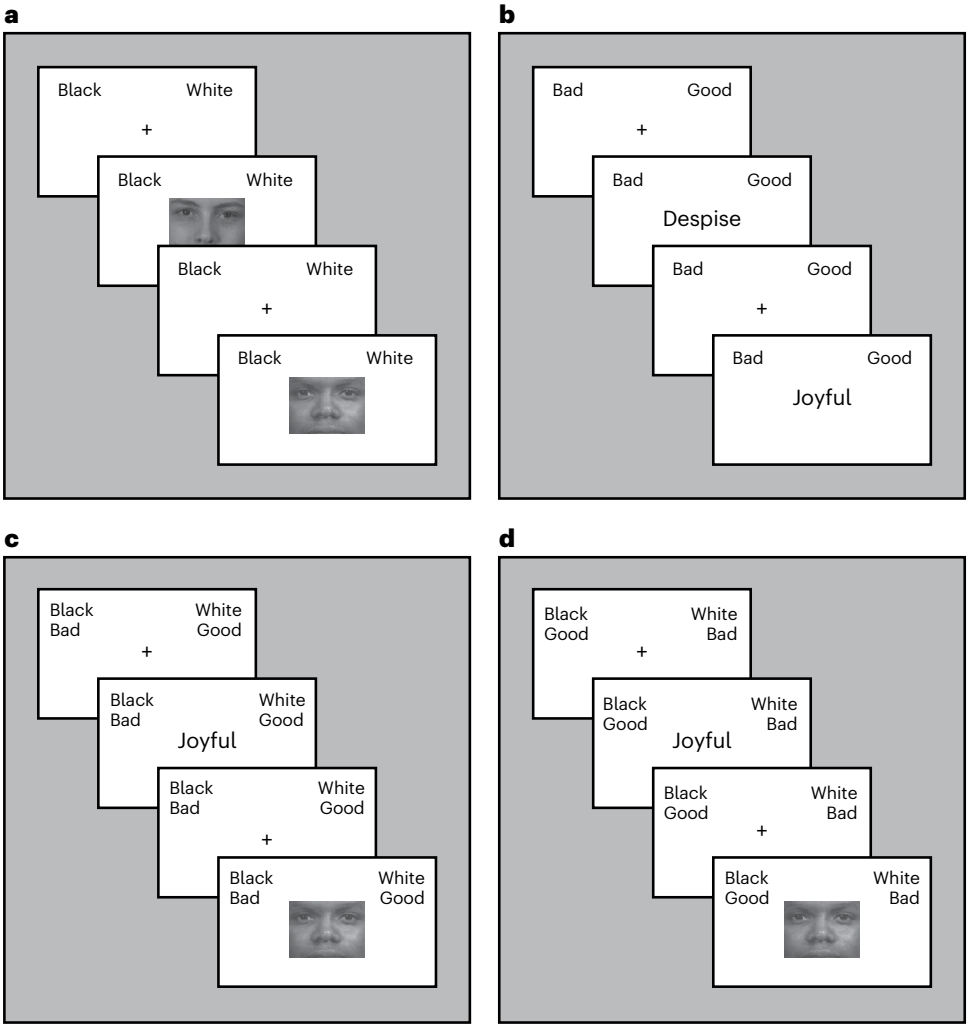

Fig. 1 | Standard IAT schematics. a–**d**, Participants classified centrally presented stimuli into categories presented in the top left and top right of their monitor screen. Concept-only blocks had participants classify concept-specific stimuli (for example, Black person, white person, gay person or straight person) (**a**). Attribute-only blocks had participants classify words conveying an attribute (for example, good, bad, true or false) (**b**). Mixed blocks had participants classify concept-specific stimuli and words into mixed categories that were either compatible (**c**) or incompatible (**d**) with biased beliefs. Images adapted from Project Implicit (https://implicit.harvard.edu).

versions of the IAT may have different construct validities: whereas some authors find that IATs measuring political biases are valid[23], others find that IATs measuring implicit self-esteem are not[24]. Some authors suggest that the IAT may be no better or even worse than explicit measures at gauging bias[25–27]. A second critique of the IAT and *D*-scores is that responses can be falsified, which can partially explain variability in *D*-scores[10–13]. This suggests that the same subjective filters that influence a person's explicit reporting of preferences may also influence their IAT performance. These criticisms suggest that further understanding of the underlying mechanisms driving variability in *D*-scores may help explain the IAT's low construct validity and the ability to falsify one's performance.

The original and most common interpretation of the IAT effect is that people process the compatible blocks more quickly due to complementary associative memory content[4], a mechanism reflecting decision ease. Incompatible blocks are responded to more slowly due to the contents of associative memory conflicting with task instructions, reflecting lesser decision ease. However, some researchers have suggested that the reason participants slow down on incompatible blocks may not entirely be due to their decision ease but also because they become more cautious of making an error on incompatible trials[5].

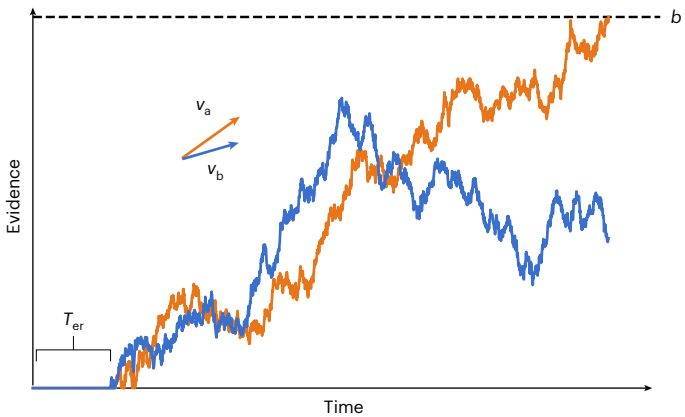

Fig. 2 | RDM schematic. Two competing accumulators race at average drift rates v_a and v_b until either reaches the evidence threshold b, at which point both processes terminate and the winning accumulator determines the chosen category. Time spent deliberating plus any non-decision time T_{er} corresponds with the choice's response time.

Table 1 | Design table

Question	Hypothesis	Sampling plan (for example, power analysis)	Analysis plan	Interpretation given to different outcomes
Do decision ease (that is, the rate of evidence accumulation), response caution (that is, the evidence threshold) and non-decision time differ between compatible and incompatible blocks?	All three processes—response caution, decision ease and non-decision time—will differ between mixed blocks for any IAT.	Our pilot analyses revealed strong evidence for effects of decision ease, response caution and non-decision time between blocks (all BF > 10 for at least moderate effects). Our exploratory analyses at the pilot stage consisted of 34,743 participants. We expected to have 109,417 participants in the final confirmatory sample and a minimum of 2,496 participants for any IAT.	We conducted three Bayesian equivalence tests for each of the 39 IATs—each one testing whether the difference in response caution, decision ease or non-decision time is different from a region of practical equivalence (ROPE) to 0 ([−0.1, 0.1] for small effects, [−0.25, 0.25] for moderate effects and [−0.4, 0.4] for large effects).	A positive difference in response bias between the two blocks indicates that the person is more cautious in the incompatible block. A negative difference between the blocks for the decision ease parameter indicates a stronger association between target–attribute pairs in the compatible block than in the incompatible block. A negative difference between the blocks for non-decision time indicates faster non-decision processes in the incompatible block than in the compatible block.
Does response caution explain significant variance in the *D*-score, above and beyond the decision ease and non-decision time scores?	(1) The response caution score will have a greater effect on the *D*-score than either decision ease or non-decision time scores. (2) The response caution score will explain significant variance in the *D*-score, above and beyond that explained by either decision ease or non-decision time scores.	The minimum sample size for any particular IAT is N=1,327. Pilot analyses suggested large effect sizes. A power analysis with a large effect size and a power of 0.95 suggested that N=238 for each IAT is sufficient for the analyses.	We conducted 39 hierarchical regression models, one for each IAT dataset, where decision ease and non-decision time scores served as the first two predictors in block 1, response caution was added as a predictor in block 2 and *D*-score served as the outcome measure. F-change tested the first hypothesis, and standardized β coefficients in block 2 tested the second hypothesis. We repeated this for all six *D*-score algorithms.	If the response caution score explains a significant portion of the variance in the *D*-score, that suggests that the *D*-score is confounded by processes unrelated to the activation of associative memory content, as is assumed in the standard definition of implicit bias.
Do response caution, decision ease and non-decision time on the IAT predict explicit preference for that IAT's topic?	The response caution score will predict explicit preference, but neither the decision ease score nor the non-decision time score will predict explicit preference.	The minimum sample size for any particular IAT is N=1,327. Pilot analyses suggested small effect sizes. A power analysis with a small effect size and a power of 0.95 suggested that N=238 for each IAT is sufficient for the analyses.	We conducted 39 ordinary least-squares regression models, one for each IAT dataset, where the decision ease score, the response caution score and the non-decision time score served as predictors of explicit preference.	If the response caution score predicts explicit preference but the decision ease score does not, that suggests that explicit preference reflects response caution more than it does processes traditionally assumed to reflect implicit bias.

Response caution reflects people slowing down on incompatible trials to avoid making mistakes. Response caution is caused either because the incompatible trials are more difficult and require slowing down[28], or because they impose pressure to fake performance[29,30].

Notice the distinction between decision ease and response caution: whereas decision ease is driven by existing associative memory between concepts and attributes, response caution is driven by subjective control to improve accuracy in performance. Previous studies have mostly neglected the question of whether these mechanisms are independently responsible for variance in IAT performance or whether they serve as a link between IAT performance and explicit preferences. One exception is work by Klauer and colleagues[31], who tested the association between IAT performance and explicit preferences in the context of political attitudes, finding that preferences were only explained by decision ease. This work, however, was limited in power[32] and topic. A comprehensive analysis of whether decision ease and response caution predict IAT effects and whether either of those predict explicit preference is needed across a wide range of IAT topics to further understand the notion of implicit bias.

In response to long-standing issues with the IAT, a growing number of researchers advocate for transitioning away from summary statistics such as *D*-scores and instead using more sophisticated computational approaches[33–35]. One such approach is the racing diffusion model (RDM)[36], a mathematical model that assumes decisions are made by accumulating bits of evidence for competing choices over time until some evidence threshold is met for either of those choices (Fig. 2). The rates of evidence accumulation are reflected in the RDM's 'drift rate' parameter, which can be thought of as decision ease. This ease is driven by the activation of associative memory content[37,38], such as the biased association between white persons and good. Conversely, the RDM's 'evidence threshold' parameter reflects response caution, which is responsible for speed–accuracy trade-offs. Since the earliest explorations of the IAT, diffusion models have been used for examining IAT performance[30,31,39,40]. However, these attempts mainly focused on examining either changes in ease and caution between blocks or the association between decision ease and the *D*-score, but not response caution as a separate predictor of the *D*-score. A third, less relevant property captured by the RDM is non-decision time, or the proportion of response time unrelated to processes of decision-making, such as stimulus encoding or pre-motor planning. Unlike decision ease or response caution, prior work suggests that non-decision time is related to neither method-specific nor construct-specific variance in IAT performance[31]. Nevertheless, accounting for variability in non-decision time can contribute to more precise estimates of decision ease and response caution.

The distinction between decision ease and response caution raises questions about the degree of conscious control people may have over either of these mechanisms. Many leading experts on decision diffusion modelling describe decision ease as reflecting abilities not under subjective control, whereas response caution is strategic and varies with participants' relative emphasis on speed versus accuracy[41–44]. In support of this description, researchers have found that only decision ease adapts to information contained by masked or suppressed stimuli outside of conscious awareness[45,46]. Response caution, in contrast, seems to be susceptible to instructional emphases on speed or accuracy[47–51] and has been linked to a network of brain regions associated with voluntary action and effortful control over behaviour[52–55]. Decision ease may also be affected by temporary response strategies that mimic shifts in mental associations, such as faking[30,56]. However, these effects probably stem from task-specific adaptations rather than genuine changes in underlying associations. Thus, decision ease may serve as a more reliable indicator of unconscious processes under

Table 2 | IAT topics and sample sizes

IAT topic	Sample with IAT	Sample with IAT and explicit preference
Age (old/new)	3,085	2,001
Businesses (corporations/non-profits)	2,832	1,804
Centuries (1950/2050)	3,044	2,435
Centuries-danger (1950/2050)	2,982	2,431
Clarity (ambiguous/clear)	3,016	2,371
Complexity (complex/simple)	2,960	2,218
Cultural ideology 1 (collective/individual)	2,992	2,271
Cultural ideology 2 (community/individual)	2,963	2,265
Cultural ideology 3 (group/individual)	3,092	2,067
Direction (backward/forward)	2,986	2,067
Economic policy (regulation/markets)	2,673	1,569
Economic systems (socialism/capitalism)	2,818	1,620
Employment (labour/management)	2,880	1,924
Equality (unequal/equal)	3,014	2,169
Explanations-truth (conspiracy/accident)	2,825	2,045
Fairness (biased/fair)	2,953	2,075
Goals (duty/hope)	3,113	2,058
Government systems (fascism/democracy)	2,823	2,027
Identity (other/self)	2,932	1,769
Justice (injustice/justice)	2,928	1,952
Likelihood (possible/certain)	2,919	1,853
Location (foreign/local)	2,942	2,025
Novelty (novel/familiar)	2,910	1,848
Order (anarchy/hierarchy)	2,829	2,191
Origins-truth (creationism/evolution)	3,055	2,224
Parenting (strict/nurturing)	3,027	1,928
Parents (father/mother)	3,142	2,074
Philosophies-truth (religion/science)	3,049	1,849
Preservation (preserve/change)	2,933	2,248
Protest (protest/accept)	2,994	1,849
Race (Black/white)	3,200	2,140
Risk (risky/cautious)	2,848	1,678
Secularity (state/church)	3,036	1,728
Sexuality (gay/straight)	3,104	1,760
Steps (restore/progress)	2,876	1,780
Strategy (attack/defend)	2,943	2,189
Threat-truth (danger/safety)	2,939	2,182
Time 1 (future/present)	2,946	2,024
Time 2 (past/present)	2,998	1,870

typical conditions, while response caution appears at least partially under conscious control.

In the present study, we use the RDM to distinguish between decision ease and response caution in predicting both *D*-scores and explicit preferences. First, we aimed to uncover the degree to which IAT effects are driven by decision ease, response caution and non-decision time. We hypothesized that decision ease, response caution and non-decision time would on average significantly differ between compatible and incompatible IAT blocks (Hypothesis 1; Table 1). Second, we hypothesized that response caution would have the greatest effect on the *D*-score, explaining variance in *D*-scores above and beyond that explained by both decision ease and non-decision time (Hypothesis 2). In addition to the traditional *D*-score, we tested this hypothesis on five alternative *D*-score algorithms[14,15]. Finally, we aimed to uncover which of those mechanisms (decision ease, response caution and non-decision time) were more associated with explicit preferences. On the basis of the majority of previous research suggesting that both response caution and explicit preferences are susceptible to subjective control, and supported by our pilot analyses with an exploratory dataset, we hypothesized that explicit preferences would be predicted solely by response caution and not by decision ease or non-decision time (Hypothesis 3). We tested these hypotheses using a large dataset of 115,601 unique IAT sessions made up of 39 different IATs.

Results

We conducted preregistered confirmatory analyses on a large held-out sample of 115,601 unique IAT sessions, collected as part of the Ideology 2.0 Study[57] by Project Implicit from December 2007 to June 2012 (see 'Pilot data' in Methods for an earlier exploratory analysis on which we based our hypotheses; see https://osf.io/e97rf for the preregistration). Thirty-nine unique IAT topics were analysed among the unique IAT sessions, each including at least 2,673 participants (mean, 2,964.128; s.d. = 103.707; Table 2). See Supplementary Information for all software used, sample demographics (Supplementary Table 3) and additional information on the dataset source and other sample considerations.

We fit both trial-level choice and response time data using the RDM for each IAT separately (see 'Analysis plan: Racing diffusion modelling' in Methods for the details). Decision ease and response caution were defined as the difference in RDM average rate of evidence accumulation and the evidence threshold, respectively, between the incompatible and compatible blocks. Bayesian equivalence tests revealed strong evidence for a large effect of decision ease between incompatible and compatible blocks (mean difference, −23.316; 95% highest density interval (HDI), (−25.972, −20.689); Bayes factor (BF), >1,000), with 38 of the 39 IATs having negative effects (Fig. 3), suggesting less ease in the incompatible block. We also observed strong evidence for a large effect of response caution between blocks (mean difference, 46.318; 95% HDI, (44.127, 48.475); BF > 1,000), with all 39 of the 39 IATs having positive effects, suggesting greater response caution in the incompatible block. Finally, we observed strong evidence for a large effect of non-decision time between blocks (mean difference, −37.239; 95% HDI, (−40.773, −34.092); BF > 1,000), with 30 of the 39 IATs having negative effects, suggesting less non-decision time in the incompatible block. See Supplementary Table 1 for individualized effects from each IAT and Supplementary Fig. 1 for visualized non-decision time.

Although decision ease, response caution and non-decision time were all significant predictors of *D*-score (absolute mean $\beta_{ease} = 0.507$; 95% confidence interval (CI), (0.487, 0.528); $z_{38} = 29.374$; two-tailed $P < 0.001$; absolute mean $\beta_{caution} = 0.809$; 95% CI, (0.788, 0.829); $z_{38} = 168.782$; two-tailed $P < 0.001$; absolute mean $\beta_{ndt} = 0.614$; 95% CI, (0.595, 0.638); $z_{38} = 55.964$; two-tailed $P < 0.001$), the absolute effect of response caution was greater on average than those of both decision ease and non-decision time (Fig. 4). We observed this for each of the six *D*-score algorithms (Fig. 5). Hierarchical regression models predicting *D*-score revealed that although a model including both decision ease and non-decision time accounted for a significant proportion of variance in individual *D*-scores (mean $R^2 = 0.627$; 95% CI, (0.616, 0.637)), including response caution as a predictor explained greater variance above and beyond decision ease and non-decision time (mean $R^2 = 0.787$; 95% CI, (0.779, 0.796)).

After establishing associations with the *D*-score, we next examined the associations between the RDM mechanisms and explicit preferences. We found that explicit preferences were predicted by decision ease, response caution and non-decision time (mean $\beta_{ease} = -0.117$;

https://doi.org/10.1038/s41562-026-02439-y

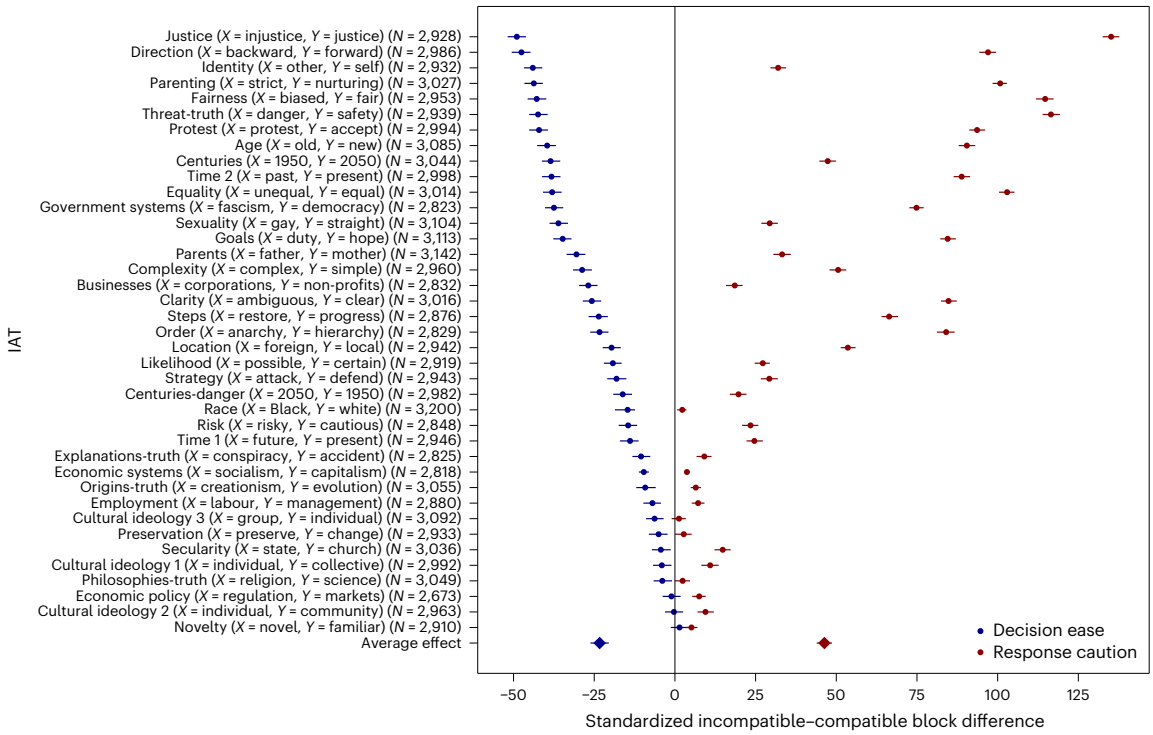

Fig. 3 | RDM parameter effects. Standardized mean differences between the posterior predictive densities of the bias-incompatible and bias-compatible blocks of each IAT (N = 115,601 participants), for decision ease (that is, the average drift rate; blue) and response caution (that is, the evidence threshold; red). Non-decision time is not shown. The diamonds indicate the average difference across IAT topics. The error bars show 95% HDIs.

95% CI, (−0.159, −0.074); $z_{38} = -9.796$; two-tailed $P < 0.001$; mean $\beta_{caution} = 0.241$; 95% CI, (0.198, 0.284); $z_{38} = 11.204$; two-tailed $P < 0.001$; mean $\beta_{ndt} = 0.184$; 95% CI, (0.149, 0.220); $z_{38} = 10.160$; two-tailed $P < 0.001$; Fig. 6) on average across IATs. However, the absolute magnitude of the effect of response caution on explicit preference was greater than the effects of both decision ease and non-decision time on explicit preference. A follow-up exploratory meta-regression supported this: response caution effect sizes were significantly larger than those of decision ease ($\beta = 0.124$; 95% CI, (0.076, 0.172); $z_{76} = 5.032$; one-tailed $P < 0.001$). Together with our hierarchical regression findings, this suggests that IAT D-scores are weak measures of implicit bias as it has been commonly defined. Response caution is largely responsible for IAT D-scores and their association with explicit preference.

Discussion

In the current project, we evaluated the influence of response caution, decision ease and non-decision time on IAT performance. To this end, we fit the RDM to IAT data curated from the Project Implicit Ideology 2.0 Study, representing 115,601 unique participants and 39 unique IAT topics. As predicted by our Hypothesis 1, we observed significant differences in RDM parameters between bias-compatible and bias-incompatible trials. Specifically, participants experienced more response caution, less decision ease and less non-decision time on bias-incompatible trials than on bias-compatible trials. Supporting our Hypothesis 2, we further observed that response caution explained significantly more variance in D-scores above and beyond both decision ease and non-decision time. This result generalized to five other variations in calculating the D-score. In Hypothesis 3, we predicted that explicit preferences would primarily relate to response caution over decision ease and non-decision time. In partial support of our hypothesis, explicit preference was predicted better by the difference between response caution in the incompatible and compatible blocks than by the difference in decision ease and the difference in non-decision time.

We draw three conclusions from the current findings, which align with those of Klauer and colleagues[31]. First, we conclude that if a researcher is interested in real-world behaviour, response caution may be the preferred measure. We found that response caution was a greater predictor of both performance on the IAT and explicit preferences, suggesting that it plays a significant role in both IAT performance and the explicit endorsement of attitudes. Caution is an important component in exercising beliefs, as other researchers have previously identified[58], and implicit bias researchers might consider a person having caution to control their biased behaviour to be a desirable trait. The D-score, however, masks this trait, as greater caution in the face of bias-incompatible trials on the IAT results in a higher D-score, confusing caution for increased bias. This conflation could explain some findings that the IAT has poor external validity[5,21,22]. Researchers who are interested in developing and testing the efficacy of bias intervention and training programmes may prefer to track changes in response caution instead of D-score as a whole[30]. Researchers may also be interested in understanding why people exert caution and the extent to which it may be conscious. Diffusion models do not themselves make claims about the motivation for response caution, and future research should investigate its function specific to the expression of bias.

Second, we conclude that if a researcher is interested in measuring implicit bias, decision ease emerges as the potentially preferred candidate. It is crucial to dissociate any measure of implicit bias from explicit measures of preference, as emphasized by most definitions of implicitness and the original purpose of the IAT[2]. Recent work suggests that the IAT D-score largely fails to provide this dissociation for lack of discriminant validity[25]; however, decision ease may serve to replace the D-score as a more valid measure. The small correlation magnitude between decision ease and explicit preference meets most rule-of-thumb criteria for discriminant validity, and although this correlation is statistically significant, our atypically large sample size paired with the low correlation magnitude demand that significance be treated with caution[59]. That said, multimethod studies and tests of convergent validity are necessary before decision ease can be endorsed with confidence as a more discriminant measure. Future research should identify decision ease with diffusion modelling across a variety

https://doi.org/10.1038/s41562-026-02439-y

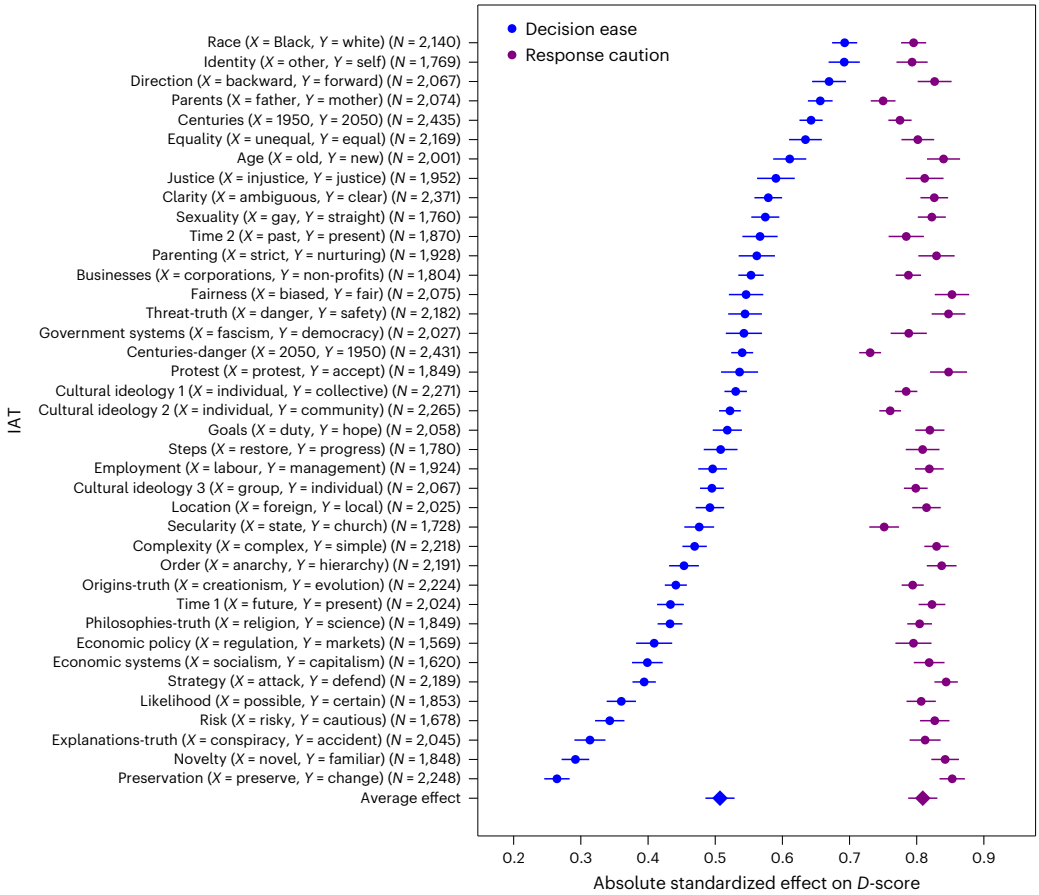

Fig. 4 | Effects of decision ease and response caution on *D*-score. Absolute standardized effects of decision ease and response caution on *D*-score ($N = 78,578$ participants). The dots represent absolute standardized β coefficients identified from regression models of *D*-score including fixed effects of decision ease, response caution and non-decision time (not shown). The diamonds indicate the average effect across IAT topics. The error bars show 95% CIs.

of implicit measures (for example, the evaluative priming task and the affective misattribution paradigm) and evaluate its validity as an implicit construct.

Third, regardless of whether a researcher is interested in real-world behaviour or implicit bias, it is clear that the *D*-score is an insufficient and flawed measure of both. It conflates response caution and decision ease, which can be considered two opposing forces. For example, a very biased person with lower decision ease on incompatible trials could achieve a lower *D*-score by exerting less caution than someone with similar bias but more caution. The *D*-score would suggest that the person who lacks caution when faced with information incompatible with their beliefs is less biased than a similar person with more caution. This is problematic for identifying bias, as it makes the *D*-score an ill-suited measure for the efficacy of implicit bias training and intervention programmes. Diffusion models solve this issue of conflation and allow both distinct behavioural components to be broken up and uniquely identified, providing a more nuanced understanding of the cognitive processes underlying biased behaviour.

We believe that these three conclusions provide much clarity on a complicated construct that has thus far suffered from limited data and analytic methods. We also believe that these complications can be assuaged by rethinking ill-defined, domain-general accounts of implicit bias with more formalized mechanistic explanations. By decomposing IAT performance into decision ease and response caution, we offer this more formalized framework for understanding and measuring the components that make up bias, some of which may be more implicit than others. Future research building on these insights will be crucial for determining whether these components truly offer a more compelling alternative to the *D*-score.

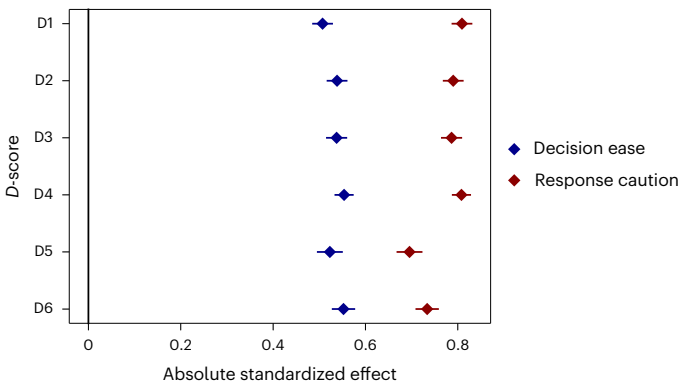

Fig. 5 | Effects of decision ease and response caution on each of the six *D*-score algorithms. Average absolute standardized effects of decision ease and response caution on each of the six *D*-score algorithms ($N = 78,578$ participants). The diamonds indicate the average standardized β coefficients identified from regression models of *D*-score including fixed effects of decision ease, response caution and non-decision time (not shown). The error bars show 95% CIs.

Limitations and future direction

Although our research provides a comprehensive reevaluation of the IAT *D*-score, it is imperative to recognize several limitations that may impact the generalizability of our findings. The first limitation is that our current results bind together a variety of types of IAT. Not all IATs elicited the same behaviour: on some IATs, such as race, decision ease was more predictive of explicit preference than was response

https://doi.org/10.1038/s41562-026-02439-y

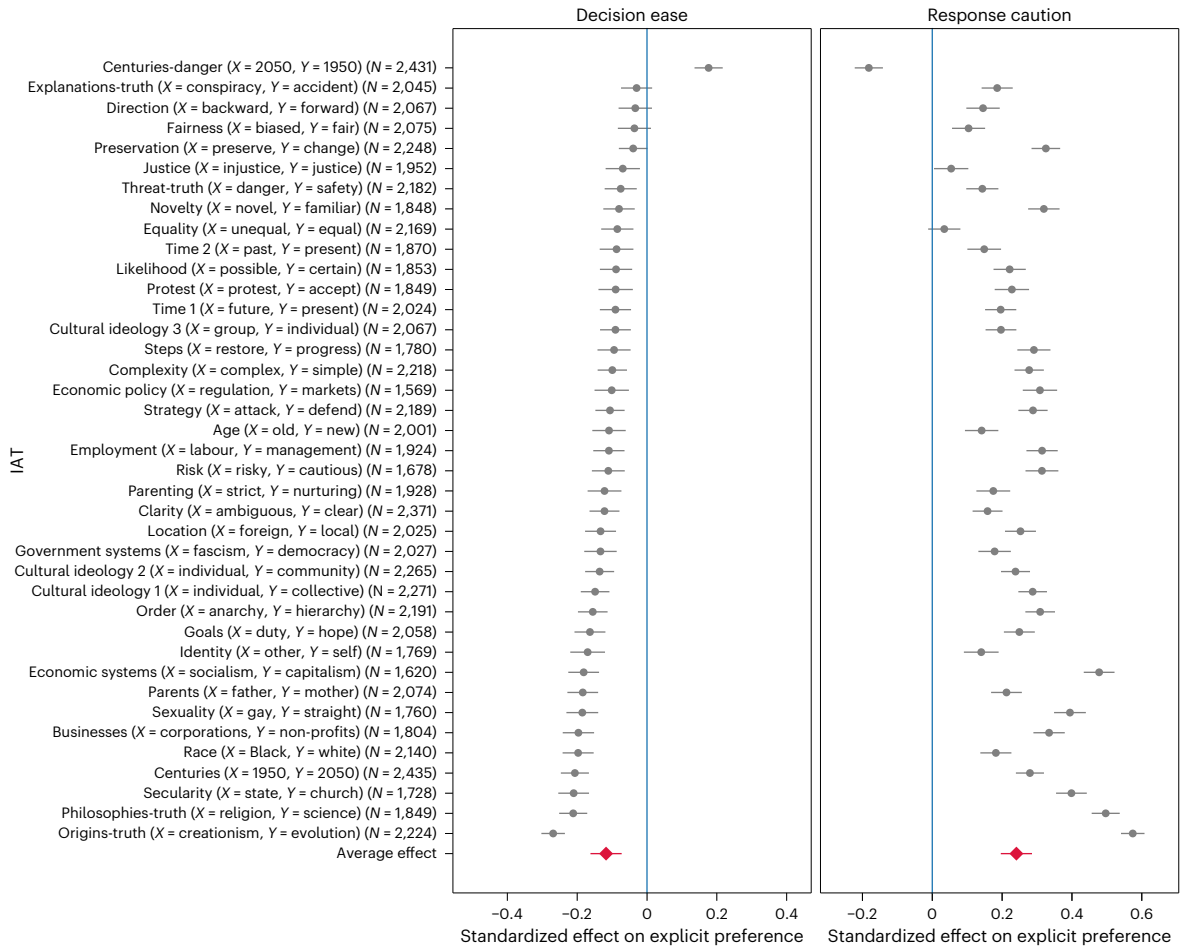

Fig. 6 | Effects of decision ease and response caution on explicit preference. Standardized effects of decision ease (left) and response caution (right) on explicit preference (N = 78,578 participants). The dots represent standardized β coefficients identified from regression models of explicit preference including fixed effects of decision ease, response caution and non-decision time (not shown). The red diamonds indicate the average effects across IAT topics. The error bars show 95% CIs.

caution. Although implicit and explicit biases are argued to be distinct constructs, Greenwald and Banaji[1] suggest that dissociability "is not a necessary condition for identifying attitudes as implicit". People can still explicitly endorse beliefs that happen to overlap with their implicit biases[60]. Future research should continue to explore specific IAT topics independently and aim to understand the conditions for overlapping implicit and explicit biases. Additionally, many of the IATs included in this analysis are under-investigated relative to others, and future work should establish their convergent validity with other implicit measures.

The second limitation is that error trials were not included in our models. The reason for this was that, during data collection, the final reaction time was recorded only when a participant made the correct choice. Intermediate, incorrect reaction times were not recorded. Because responding accurately on the IAT involves overcoming a conflict between responding on the basis of category membership and responding on the basis of biased associations, including data from error trials could have provided additional information for evaluating response caution and decision ease. Future research should assess these parameters while considering inaccurate responses.

A third limitation of the current design is that decision ease, although identified as a potentially better measure of implicit bias, may also reflect phenomena unrelated to implicit biases, such as task switching. Task switching has been found to affect decision ease on the bias-incompatible block via proactive interference[61] and is an important feature of IAT performance[62]. To mitigate the effect of task switching, the standard IATs used here included 40 practice trials between the 60 trials in each of two mixed-configuration blocks—compatible and incompatible—thereby reducing proactive interference. Nevertheless, there is still a possibility that changes in decision ease between blocks may be due in part to some proactive interference brought on by task switching, despite the efforts to mitigate it.

Efforts to mitigate task-switching effects may also explain our observed changes in non-decision time. Non-decision time was observed to be notably shorter in incompatible blocks than in compatible blocks. This conflicts with Klauer and colleagues' observation that non-decision time is longer in incompatible blocks. There are two key reasons that may account for this discrepancy. First, Klauer and colleagues acknowledge that changes in non-decision time most likely reflect proactive interference between mixed-configuration blocks. Our IATs included a greater proportion of practice trials relative to mixed-configuration trials between the mixed-configuration blocks than did Klauer's IATs, which should further reduce proactive interference and potentially lead to shorter non-decision times in incompatible blocks. Second, recent reports indicate that when simulated non-decision time remains unchanged between blocks, spurious negative correlations between response caution and non-decision time may arise[63]. We believe that the negative difference in non-decision time in our analysis may reflect such a spurious correlation with the observed positive difference in response caution. Importantly, when we explicitly fixed non-decision time to be equal across blocks in an alternative model specification, we still observed a positive difference in response caution. This confirms that the change in response caution

is real and not merely an artefact of spurious correlation. However, with our expanded models, we caution against overinterpreting the reduction in non-decision time itself, as it may primarily reflect the interplay of response caution with methodological differences rather than a meaningful process-level change.

A fourth limitation of the current work is its reliance on a traditional interpretation of implicit bias, which frames it as unconscious and dissociable from conscious explicit preference. While this perspective has been foundational to our work and that of other IAT researchers, it may not fully capture the complexity of implicit bias. For instance, Krajbich suggests a more pragmatic approach, defining implicit bias as an unintentional and undesirable contamination of behaviour, rather than a strictly unconscious process[64]. This view also raises the possibility that response caution, considered separate from implicit processes, may play a more significant role in implicit bias that previously recognized. Incorporating these contemporary perspectives in future research could provide a more nuanced understanding of implicit bias.

Overall, our findings provide a case for using diffusion models in implicit bias research. Our comprehensive study of 39 IAT topics across 115,601 participants suggests that IAT *D*-scores capture a good candidate for implicit bias—decision ease—but diffusion models must be used to identify it. Once it has been identified, we can dissociate decision ease from other processes such as response caution and use it as the preferred marker for implicit bias. Future research should adopt diffusion modelling to reevaluate interpretations of implicit bias research that relies on *D*-scores. This approach could resolve long-standing issues with IAT construct validity; future IAT validation studies should incorporate processes of diffusion to determine whether decision ease can serve as a more valid replacement of the *D*-score. Finding a preferred measure of implicit bias would facilitate more ecologically valid research, such as that investigating the prognostic value of the IAT on actual behaviour and the potential value of implicit bias training.

Methods

Ethics information

Our research complies with all ethical regulations. Our analyses use data collected from the Ideology 2.0 Study[57] by Project Implicit from December 2007 to June 2012. Data collection was approved by the University of Virginia's Institutional Review Board (IRB) for Social and Behavioral Sciences, and informed consent was obtained from all participants. The Case Western Reserve University IRB determined that further IRB approval was not required to conduct secondary analyses on these data.

Pilot data

We conducted exploratory analyses on a subsample of the data for proof of concept, separate from a larger held-out confirmatory sample. These analyses were the bases of our expected effect sizes and confirmatory hypotheses. Of the 39 IATs we analysed, each included at least 746 participants (mean, 818.589; s.d. = 30.714). We fit both trial-level choice and response time data using the RDM for each IAT separately (see 'Analysis plan' for the details). Decision ease, response caution and non-decision time were defined as the difference in RDM average rate of evidence accumulation, the evidence threshold and non-decision time, respectively, between the incompatible and compatible blocks. Bayesian equivalence tests revealed strong evidence for a large effect of decision ease between incompatible and compatible blocks (mean difference, −5.738; 95% HDI, (−8.471, −3.008); BF > 1,000), with 30 of the 39 IATs having negative effects, suggesting less ease in the incompatible block. We also observed strong evidence for a large effect of response caution between blocks (mean difference, 39.479; 95% HDI, (36.803, 42.252); BF > 1,000), with 37 of the 39 IATs having positive effects, suggesting greater response caution in the incompatible block. Finally, we observed strong evidence for a large effect of non-decision time between blocks (mean difference, −31.538; 95% HDI, (−34.305, −28.894); BF > 1,000), with 36 of the 39 IATs having negative effects, suggesting less non-decision time in the incompatible block.

Although decision ease, response caution and non-decision time were all significant predictors of *D*-score (absolute mean $\beta_{ease} = 0.353$; 95% CI, (0.236, 0.470); absolute mean $\beta_{caution} = 0.965$; 95% CI, (0.827, 1.104); absolute mean $\beta_{ndt} = 0.470$; 95% CI, (0.333, 0.607)), the absolute effect of response caution was greater on average than those of both decision ease and non-decision time. We observed this for each of the six *D*-score algorithms. Hierarchical regression models predicting *D*-score revealed that although a model including both decision ease and non-decision time accounted for a significant proportion of variance in individual *D*-scores (mean $R^2 = 0.633$; 95% CI, (0.569, 0.696)), including response caution as a predictor explained greater variance above and beyond decision ease and non-decision time (mean $R^2 = 0.784$; 95% CI, (0.742, 0.825)).

After establishing associations with the *D*-score, we next examined the associations between the RDM mechanisms and explicit preferences. We found that explicit preferences were predicted by neither decision ease nor response caution nor non-decision time (mean $\beta_{ease} = -0.045$; 95% CI, (−0.296, 0.197); mean $\beta_{caution} = 0.265$; 95% CI, (−0.021, 0.551); mean $\beta_{ndt} = 0.119$; 95% CI, (−0.162, 0.401)) on average across IATs. However, at the individual level, we found that response caution did have a substantial effect on explicit preference for 20 of the 39 IATs, whereas decision ease and non-decision time had effects on explicit preference for only 3 of the 39 IATs each. Together with our hierarchical regression findings, this suggests that IAT *D*-scores are weak measures of implicit bias as it has been commonly defined. Response caution is largely responsible for IAT *D*-scores and their association with explicit preference. We thus proposed that, should our confirmatory analyses support this, we would encourage IAT researchers to remember the literature linking response caution and conscious control when considering *D*-scores. As an alternative to *D*-scores, we would suggest that researchers use diffusion modelling to disentangle response caution from other mechanisms that they may have a greater interest in measuring, such as decision ease.

Design

Participants were randomly and blindly assigned to one of two study designs: Design A or Design B. Participants in Design A completed one standard or IAT measure of evaluation (good or bad), stereotyping identification or validation (true or false). Participants in Design B were randomly assigned to and completed two IATs. The topic of each IAT was randomly selected from a pool of 39 topics (Table 2). The participants also completed explicit preference thermometers and individual difference scales and items following the IAT(s). About 50% of participants who were given any one of the 39 IAT topics also have explicit preference data for that topic. For the confirmatory analysis, we limited our variables of interest to (1) IAT performance (choices and response times) and (2) explicit preference for the concepts targeted by the IAT. Likewise, only these two variables were considered in our pilot analysis. All other explicit measures and individual difference scales/items were to be restricted to exploratory analyses. See below for descriptions of the IAT and explicit measures.

IATs. Standard IATs were composed of four blocks, the first two of which were 20-trial single categorization practice blocks specific to either the target concept (for example, Black or white) or the attribute for association (for example, good or bad). Labels for these concepts/attributes were presented at the top left and top right of the participant's computer screen, whereas the stimuli targeted for classification were presented in the screen's centre. The participants were instructed to use the 'E' and 'I' keys on their keyboard to classify stimuli with the top-left and top-right labels, respectively. After these practice blocks, the participants then completed a 60-trial mixed

categorization block where concept and attribute labels were combined (for example, Black/bad or white/good). Another 40-trial target concept categorization block was then administered, followed by a 60-trial mixed categorization block with label combinations inverted from the previous mixed block (for example, white/bad or Black/good). See Fig. 1 for the standard IAT schematics. Task instructions and stimuli particular to each of the 39 possible IATs are available at https://osf.io/2483h/. Choices and response times were recorded for each trial. On error trials, the original reaction time was not recorded. Instead, the task persisted until the choice was corrected, and only the final reaction time for the correct choice was recorded. As a result, we cannot provide reaction time analysis for any error trials using the RDM. An average of 528,451.795 trials (s.d. = 24,651.212) of (only) correct responses were included from each of the 39 IATs for analysis, with an average of 178.217 trials (s.d. = 2.828) per participant after exclusions. See 'Analysis Plan: Pre-Processing – Exclusions' for more information on exclusion criteria.

Explicit preference thermometers. Following the IAT(s), participants were given a one-item relative preference (or liking) thermometer measure. The topic of this measure was matched to the topic of the IAT performed in Design A (one IAT test) and at random to that of one of the IATs performed in Design B (two sequential IAT tests). Explicit preference was determined either as the relative preference for one target concept over another (for example, Black over white on a scale from 1 to 7, with 1 being a strong preference for Black over white and 7 being a strong preference for white over Black) or as the likeability of one specific target (for example, white on a scale from 1 to 7, with 1 being strong unlikability of white and 7 being strong likability of white). Each participant was given either the relative preference measure or the likability measure.

Sampling plan

Participants. We tested our hypotheses on a confirmatory held-out subsample of the Ideology 2.0 dataset. Within this dataset, we analysed data from 115,601 unique sessions, spread across 39 IAT topics. On average, 2,964.128 (s.d. = 103.707) participants completed each IAT, 2,014.821 (s.d. = 215.662) of whom also completed an explicit preference measure for that IAT's topic. See Table 2 for sample size by IAT topic. This held-out sample was masked at the time of submitting the Stage 1 Registered Report and preregistered.

Expected effect sizes. Effect sizes were informed using our pilot analyses of the exploratory dataset. Our preregistered Hypothesis 1—that response caution and decision ease would differ between incompatible and compatible blocks—was informed by a large effect of block on both response caution and non-decision time (BFs > 1,000) and a moderate effect of block on decision ease (BF = 11.97). Preregistered Hypothesis 2—that response caution would explain significant variance in *D*-scores and have a greater effect on the *D*-score than decision ease—was supported by a significantly larger effect of response caution than either decision ease or non-decision time on the *D*-score, as well as significant variance explained in *D*-scores by response caution above and beyond ease and non-decision time. We anticipated a large effect of response caution and moderate effects of decision ease and non-decision time on the *D*-score. Preregistered Hypothesis 3—that response caution would predict explicit preference whereas decision ease and non-decision time would not—was supported by a small effect of response caution on explicit preference, and no effect of decision ease or non-decision time on explicit preference.

Power analysis. For preregistered Hypotheses 2 and 3, a power analysis with a small effect size of 0.2 and power at 0.95 suggested that $N = 237$ would be sufficient to detect effects. In terms of Hypothesis 2, although we expected large effect sizes, our sample size was sufficient to detect a smaller effect.

Analysis plan

Pre-processing. *Exclusions.* Error trials were excluded from model fitting and analysis because participants were instructed to correct their mistakes and only the final reaction time after correction was recorded (as opposed to one reaction time after error and another reaction time after correction). As such, response times recorded for incorrect trials were confounded by multiple decision processes, and their reaction times were unsuitable for RDM modelling. Furthermore, the proportion of error trials on IATs is often too small to reliably compare correct and error reaction time distributions[64]. Across the 39 IATs included in our pilot analyses, an average 8.24% (s.d. = 1.12%) of observations were error trials. We also excluded any trials with reaction times less than 200 ms, as these are likely to be false starts, and greater than 5,000 ms, as these are likely to be responses while the participant was distracted[40,65]. Furthermore, our RDM models require variability in sorting patterns to estimate both drift rates, and therefore any participants who exclusively pressed only one response key in any block were excluded.

Although all participants with IAT data that met the inclusion criteria were used for RDM model fitting, only those with IAT and explicit preference data were modelled at the participant level. Group-level hyperparameters were fitted with all data that passed the inclusion criteria. Those hyperparameters were used as priors for estimating participant-level parameters. However, we estimated participant-level parameters only for those participants with IAT and explicit preference data.

Outliers. We did not exclude any trials or participants on the basis of outlier detection.

Racing diffusion modelling. IAT choices and response times were modelled as racing diffusion processes. The RDM assumes that peoples' responses are functions of competing evidence accumulators that each integrate noisy information over time. A choice is made once a choice's associated accumulator surpasses an evidence threshold. Our models include two competing accumulators per IAT block, each representing a noisy evidence accumulation process towards one of the two categorization options.

RDMs (Fig. 2) decompose choices and response times into four latent parameters—an evidence threshold or boundary (b), average drift rates for each accumulator a (v_a), non-decision time for encoding and responding (T_{er}) and between-trial variability in accumulator starting point (kept constant in models not assuming between-trial variability; see below). Sequential sampling models such as the RDM can be used for explaining differences in response caution as shifts in evidence threshold. Greater thresholds require that a person accumulates more evidence before a decision can be made and therefore increases the likelihood of accuracy at the cost of time. Similarly, differences in decision ease or processing efficiency can be explained as differences in drift rates between IAT blocks. Greater average drift rates on white/good, Black/bad trials than on Black/good, white/bad trials are indicative of greater decision ease for white/good and Black/bad concept–attribute pairs. This corresponds to greater relative accuracy on those trials and faster response time. The IAT *D*-score, which also relies on response time, conflates these two sources of variance—response caution and decision ease. When performing the IAT, participants are asked to respond as quickly and as accurately as possible. These instructions can themselves generate a conflict between speed and accuracy, and, given that both criteria seem to be equal, participants may feel comfortable sacrificing one for the other.

RDMs offer a few major advantages over other more standard sequential sampling models for our analyses, such as classic drift diffusion models (DDMs) and linear ballistic accumulators (LBAs). First, separate accumulators allowed us to measure the magnitude of drift rate for each concept–attribute pair, unlike the relative difference that DDMs' single drift rate provides. This is important when modelling

evidence accumulation for competing classes (for example, Black/bad or white/good), as a single, relative drift rate between them could be indistinguishable between blocks while having very different drift magnitudes. For example, say a person accumulates evidence at rates of $v_a = 2$ and $v_b = 3$ on compatible blocks, and $v_a = 0.5$ and $v_b = 1.5$ on incompatible blocks. A single, relative drift rate would accumulate toward category b on both of these blocks ($v = -1$), but the difference in magnitude would be lost. A single accumulator would be blind to the average difference in evidence accumulation between blocks. Second, the RDM maintains that evidence accumulation is a noisy diffusion process, which better reflects how conflicting associations are theoretically activated than does the noiseless linear function assumed by LBAs. This noise also explains variability in response time, overcoming LBAs' dependence on between-trial variability in starting point and drift rate. LBAs assume a constant, noiseless rate of evidence accumulation, so without between-trial variability they would predict identical, non-changing response times for every trial. Third, unlike the DDM and LBA, recent work suggests that assumptions of between-trial variability are not necessary for the RDM to model systematic variability in response time, and therefore the accumulator starting bias A can be dropped for a more parsimonious account[36]. The inclusion of between-trial variability parameters in diffusion models has largely been motivated by the need to explain variation in response time, but those parameters do not actually have a role in the process model itself and do not serve good theory. The RDM better explains choices and response times without having to resort to theory-inconsistent parameters. Together, these advantages of RDMs combine the strengths of DDMs and LBAs to better model response caution and decision ease on IATs.

Evidence threshold, drift rates for each classification option and non-decision time were modelled hierarchically. We leveraged the full available dataset for each IAT to estimate group-level means and variances. These hyperparameters for evidence threshold, both drift rates and non-decision time were free to vary between IAT blocks. Hyperpriors were weakly informative and constrained to be greater than 0, with the following distributions:

$$b_\mu \sim \mathrm{N}(0.5, 1)$$

$$b_\sigma \sim \Gamma(1, 1)$$

$$v_{a,\mu} \sim \mathrm{N}(2, 1)$$

$$v_{a,\sigma} \sim \Gamma(1, 1)$$

$$T_\mu \sim \mathrm{N}(0.5, 0.5)$$

$$T_\sigma \sim \Gamma(1, 1)$$

Although the group-level parameters leveraged the full dataset available for each IAT, participant-level parameters were only estimated for participants who also completed an explicit preference thermometer for the IAT topic. Participant-level evidence thresholds, drift rates and non-decision times were free to vary with IAT block, and priors were normally distributed with location and scale set to the block-specific group-level hyperparameter. Prior work suggests that non-decision time is related to neither method-specific nor construct-specific variance in IAT performance[31] and that non-decision time can be spuriously and negatively correlated with evidence thresholds, contributing to problems interpreting either parameter when both are free to vary[63]. For these reasons, we also tested a model in which participant-level non-decision time was fixed across IAT blocks. Both models from our pilot analyses yielded similar conclusions. Therefore, because the varying model may allow for a more comprehensive understanding of how IAT decisions are made, we report analyses from the varying model here (see Supplementary Table 2 for Hypothesis 1 results using a simpler model with fixed non-decision time). All participant-level parameters were constrained to be greater than 0. The likelihood of the response time data t given the RDM parameters b, v_i and T_{er} was calculated using separate Wald distributions[66] for each competing classification i. The Wald probability density function is:

$$p(t|b, v_i, T_{er}) = \frac{b}{\sqrt{2\pi(t - T_{er})^3}} e^{\frac{-(v_i(t - T_{er}) - b)^2}{2(t - T_{er})}}$$

RDMs were separately fit to data from each of the IATs. An average of 528,451.795 trials (s.d. = 24,651.212) were included from each of the 39 IATs for RDM model fitting, with an average of 178.217 trials (s.d. = 2.828) per participant for participant-level estimates. Prior work has found that both group-level and participant-level RDM parameters can be reliably recovered from far fewer trials and participants[36]. Posterior parameter distributions were inferred using Hamiltonian Monte Carlo No-U-Turn sampling in Stan[67]. Hamiltonian Monte Carlo is a Markov chain Monte Carlo sampling method for estimating the joint posterior of our RDM parameters. Four chains were run in parallel at 2,000 iterations each. We discarded the first 1,000 iterations as warm-up samples. Chain convergence was qualified with Gelman–Rubin diagnostics $\hat{R}$ less than or equal to 1.01 (ref. 68).

Response caution score, decision ease score and non-decision time score. Following scoring methods for diffusion models used in prior IAT studies[31], we define 'response caution score' as the difference in response caution between the incompatible and compatible blocks of each IAT. Similarly, we define 'decision ease score' as the difference in estimated decision ease between the incompatible and compatible blocks of each IAT. Last, we define 'non-decision time score' as the difference in estimated non-decision time between the incompatible and compatible blocks of each IAT.

Confirmatory analyses. *Preregistered Hypothesis 1.* To test whether response caution, decision ease and non-decision time differ between the incompatible and compatible IAT blocks, we conducted 39 Bayesian equivalence tests, three for each IAT dataset—one testing whether the response caution score is different from 0, another testing whether the decision ease score is different from 0 and the last testing whether the non-decision time score is different from 0. The scores were derived from each IAT's block-specific group-level evidence threshold, group-level average drift rate and group-level non-decision time, respectively. We hypothesized that all three scores would differ between mixed blocks for any IAT. Comprehensive results from these individualized tests are reported in Supplementary Tables 1 and 2, and summary tests of aggregate effects across IATs are reported in the main text.

Bayesian equivalence tests determine whether a null value is among the credible values of the posterior distribution differences in response caution or decision ease. For each difference distribution, we established a 95% HDI representing the 95% most credible difference values. We further established a ROPE around the null difference value, specified as half of Cohen's conventional effect sizes[69], the range of −0.1 to 0.1 for small effects, −0.25 to 0.25 for moderate effects and −0.4 to 0.4 for large effects, as suggested by Kruschke and Liddell[70]. In place of null-point BFs, we used BFs versus ROPE to test interval null hypotheses. BFs versus ROPE test for the odds of a posterior distribution falling within a ROPE to a null value, relative to the odds of a prior[71]. For hypothesis testing, we specified our posterior as the standardized difference distribution and our prior as a standard Cauchy distribution. Using a BF-versus-ROPE decision rule, we accepted that a parameter

difference was practically equivalent to the null value if there was 10× ROPE odds for the posterior over the prior (BF ≥ 10) and rejected equivalence to the null value if there was 10× ROPE odds for the prior over the posterior (BF ≤ 0.1). All other cases (10 > BF > 0.1) were considered weakly informative and left to interpretation.

Preregistered Hypothesis 2. To test whether the three RDM scores predict IAT *D*-score, we conducted 39 hierarchical ordinary least-squares regression models, one for each IAT dataset. Decision ease scores and non-decision time scores were entered first, then the response caution scores, to measure the unique contribution of response caution scores to the *D*-score. For this analysis, response caution, decision ease and non-decision time scores were derived from each IAT's block-specific participant-level evidence threshold, participant-level average drift rate and participant-level non-decision time, respectively. *F*-tests were conducted between hierarchical regression blocks to determine whether there was a significant change in variance explained by including response caution score as a predictor. We hypothesized that the response caution score would explain a significant proportion of variance in *D*-scores above and beyond the decision ease score and the non-decision time score for each IAT. We further hypothesized that the response caution score would have a significantly greater effect on the *D*-score than would the decision ease score or the non-decision time score. We repeated this analysis for each of the six *D*-score algorithms.

Preregistered Hypothesis 3. To test whether response caution scores and decision ease scores predict IAT explicit preferences, we conducted 39 ordinary least-squares regression models, one for each IAT dataset, where decision ease score, response caution score and non-decision time score served as predictors of explicit preference. For this analysis, response caution, decision ease and non-decision time scores were derived from each IAT's block-specific participant-level evidence threshold, participant-level average drift rate and participant-level non-decision time, respectively. We hypothesized that the response caution score would predict explicit preference, but neither the decision ease score nor the non-decision time score would predict explicit preference.

Protocol registration

The Stage 1 protocol, as accepted by the journal on 8 November 2023, can be found at ref. 72.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data and materials are available via the Open Science Framework at https://osf.io/2r3zd/.

Code availability

All code is available via the Open Science Framework at https://osf.io/2r3zd/, under the GitHub tab. This can also be found directly via GitHub at https://github.com/GoldenbergLab/model-iat-drm-test-mechanisms-kyle.

References

1. Greenwald, A. G. & Banaji, M. R. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychol. Rev.* **102**, 4–27 (1995).
2. Greenwald, A. G. et al. Best research practices for using the Implicit Association Test. *Behav. Res.* **54**, 1161–1180 (2022).
3. Banaji, M. R. & Greenwald, A. G. *Blindspot: Hidden Biases of Good People* (Bantam Books, 2016).
4. Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. K. Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol.* **74**, 1464–1480 (1998).
5. Nosek, B. A., Greenwald, A. G. & Banaji, M. R. The implicit association test at age 7: a methodological and conceptual review. In *Social psychology and the unconscious: The automaticity of higher mental processes* (ed. Bargh, J. A.) 265–292 (Psychology Press, 2007).
6. Prestwich, A., Kenworthy, J. B., Wilson, M. & Kwan-Tat, N. Differential relations between two types of contact and implicit and explicit racial attitudes. *Br. J. Soc. Psychol.* **47**, 575–588 (2008).
7. Fazio, R. H. & Olson, M. A. Implicit measures in social cognition research: their meaning and use. *Annu. Rev. Psychol.* **54**, 297–327 (2003).
8. King, M. F. & Bruner, G. C. Social desirability bias: a neglected aspect of validity testing. *Psychol. Mark.* **17**, 79–103 (2000).
9. Gawronski, B. & Hahn, A. Implicit measures: procedures, use, and interpretation. In *Measurement in Social Psychology* (eds Blanton, H. et al.) 29–55 (Routledge, 2019).
10. De Houwer, J. Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learn. Motiv.* **37**, 176–187 (2006).
11. Röhner, J., Schröder-Abé, M. & Schütz, A. Exaggeration is harder than understatement, but practice makes perfect! Faking success in the IAT. *Exp. Psychol.* **58**, 464–472 (2011).
12. Schindler, S., Wolff, W., Kissler, J. M. & Brand, R. Cerebral correlates of faking: evidence from a brief implicit association test on doping attitudes. *Front. Behav. Neurosci.* **9**, 139 (2015).
13. De Houwer, J., Beckers, T. & Moors, A. Novel attitudes can be faked on the Implicit Association Test. *J. Exp. Soc. Psychol.* **43**, 972–978 (2007).
14. Greenwald, A. G., Nosek, B. A. & Banaji, M. R. Understanding and using the implicit association test: I. An improved scoring algorithm. *J. Pers. Soc. Psychol.* **85**, 197–216 (2003).
15. Röhner, J. & Thoss, P. J. A tutorial on how to compute traditional IAT effects with R. *Quant. Methods Psychol.* **15**, 134–147 (2019).
16. Hall, W. J. et al. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. *Am. J. Public Health* **105**, e60–e76 (2015).
17. Chapman, E. N., Kaatz, A. & Carnes, M. Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. *J. Gen. Intern. Med.* **28**, 1504–1510 (2013).
18. Hehman, E., Flake, J. K. & Calanchini, J. Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Soc. Psychol. Pers. Sci.* **9**, 393–401 (2018).
19. Van Den Bergh, L., Denessen, E., Hornstra, L., Voeten, M. & Holland, R. W. The implicit prejudiced attitudes of teachers: relations to teacher expectations and the ethnic achievement gap. *Am. Educ. Res. J.* **47**, 497–527 (2010).
20. Chin, M. J., Quinn, D. M., Dhaliwal, T. K. & Lovison, V. S. Bias in the air: a nationwide exploration of teachers' implicit racial attitudes, aggregate bias, and student outcomes. *Educ. Res.* **49**, 566–578 (2020).
21. Andersen, J. P., Di Nota, P. M., Boychuk, E. C., Schimmack, U. & Collins, P. I. Racial bias and lethal force errors among Canadian police officers. *Can. J. Behav. Sci.* **55**, 130–141 (2023).
22. Payne, B. K., Vuletich, H. A. & Lundberg, K. B. The bias of crowds: how implicit bias bridges personal and systemic prejudice. *Psychol. Inq.* **28**, 233–248 (2017).
23. Bar-Anan, Y. & Vianello, M. A multi-method multi-trait test of the dual-attitude perspective. *J. Exp. Psychol. Gen.* **147**, 1264–1272 (2018).

24. Falk, C. F., Heine, S. J., Takemura, K., Zhang, C. X. J. & Hsu, C.-W. Are implicit self-esteem measures valid for assessing individual and cultural differences? Implicit self-esteem. *J. Pers.* **83**, 56–68 (2015).
25. Schimmack, U. The Implicit Association Test: a method in search of a construct. *Perspect. Psychol. Sci.* **16**, 396–414 (2021).
26. Granados Samayoa, J. A. & Fazio, R. H. Who starts the wave? Let's not forget the role of the individual. *Psychol. Inq.* **28**, 273–277 (2017).
27. Corneille, O. & Gawronski, B. Self-reports are better measurement instruments than implicit measures. *Nat. Rev. Psychol.* **3**, 835–846 (2024).
28. Schmitz, F. & Voss, A. Decomposing task-switching costs with the diffusion model. *J. Exp. Psychol. Hum. Percept. Perform.* **38**, 222–250 (2012).
29. Röhner, J. & Thoss, P. EZ: an easy way to conduct a more fine-grained analysis of faked and nonfaked Implicit Association Test (IAT) data. *Quant. Methods Psychol.* **14**, 17–37 (2018).
30. Röhner, J. & Lai, C. K. A diffusion model approach for understanding the impact of 17 interventions on the race Implicit Association Test. *Pers. Soc. Psychol. Bull.* **47**, 1374–1389 (2021).
31. Klauer, K. C., Voss, A., Schmitz, F. & Teige-Mocigemba, S. Process components of the Implicit Association Test: a diffusion-model analysis. *J. Pers. Soc. Psychol.* **93**, 353–368 (2007).
32. Lerche, V., Voss, A. & Nagler, M. How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behav. Res.* **49**, 513–537 (2017).
33. Haines, N., Kvam, P. D., Irving, L., Smith, C. T., Beauchaine, T. P., Pitt, M. A., Ahn, W. H., & Turner, B. M. A tutorial on using generative models to advance psychological science: Lessons from the reliability paradox. *Psychol. Methods* https://doi.org/10.1037/met0000674 (2025).
34. Kvam, P. D., Irving, L. H., Sokratous, K. & Smith, C. Improving the reliability and validity of the IAT with a dynamic model driven by similarity. *Behav. Res.* **56**, 2158–2193 (2024).
35. Elder, J., Wilson, L. & Calanchini, J. Estimating the reliability and stability of cognitive processes contributing to responses on the Implicit Association Test. *Pers. Soc. Psychol. Bull.* **50**, 1451–1470 (2023).
36. Tillman, G., Van Zandt, T. & Logan, G. D. Sequential sampling models without random between-trial variability: the racing diffusion model of speeded decision making. *Psychon. Bull. Rev.* **27**, 911–936 (2020).
37. Ratcliff, R. A theory of memory retrieval. *Psychol. Rev.* **85**, 59–108 (1978).
38. Ratcliff, R., Smith, P. L., Brown, S. D. & McKoon, G. Diffusion decision model: current issues and history. *Trends Cogn. Sci.* **20**, 260–281 (2016).
39. Brendl, C. M., Markman, A. B. & Messner, C. How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *J. Pers. Soc. Psychol.* **81**, 760–773 (2001).
40. Röhner, J. & Ewers, T. Trying to separate the wheat from the chaff: construct- and faking-related variance on the Implicit Association Test (IAT). *Behav. Res.* **48**, 243–258 (2016).
41. Wagenmakers, E.-J. Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *Eur. J. Cogn. Psychol.* **21**, 641–671 (2009).
42. Shevlin, B. R. K., Smith, S. M., Hausfeld, J. & Krajbich, I. High-value decisions are fast and accurate, inconsistent with diminishing value sensitivity. *Proc. Natl Acad. Sci. USA* **119**, e2101508119 (2022).
43. Pleskac, T. J., Cesario, J. & Johnson, D. J. How race affects evidence accumulation during the decision to shoot. *Psychon. Bull. Rev.* **25**, 1301–1330 (2018).
44. Forstmann, B. U., Ratcliff, R. & Wagenmakers, E.-J. Sequential sampling models in cognitive neuroscience: advantages, applications, and extensions. *Annu. Rev. Psychol.* **67**, 641–666 (2016).
45. Barbosa, L. S., Vlassova, A. & Kouider, S. Prior expectations modulate unconscious evidence accumulation. *Conscious. Cogn.* **51**, 236–242 (2017).
46. Vlassova, A., Donkin, C. & Pearson, J. Unconscious information changes decision accuracy but not confidence. *Proc. Natl Acad. Sci. USA* **111**, 16214–16218 (2014).
47. Ratcliff, R. & Rouder, J. N. A diffusion model account of masking in two-choice letter identification. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 127–140 (2000).
48. Katsimpokis, D., Hawkins, G. E. & Van Maanen, L. Not all speed–accuracy trade-off manipulations have the same psychological effect. *Comput. Brain Behav.* **3**, 252–268 (2020).
49. Milosavljevic, M., Malmaud, J., Huth, A., Koch, C. & Rangel, A. The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgm. Decis. Mak.* **5**, 437–449 (2010).
50. Voss, A., Rothermund, K. & Voss, J. Interpreting the parameters of the diffusion model: an empirical validation. *Mem. Cogn.* **32**, 1206–1220 (2004).
51. Matzke, D. & Wagenmakers, E.-J. Psychological interpretation of the ex-Gaussian and shifted Wald parameters: a diffusion model analysis. *Psychon. Bull. Rev.* **16**, 798–817 (2009).
52. Cavanagh, J. F. et al. Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nat. Neurosci.* **14**, 1462–1467 (2011).
53. Frank, M. J. et al. fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *J. Neurosci.* **35**, 485–494 (2015).
54. Van Maanen, L. et al. Neural correlates of trial-to-trial fluctuations in response caution. *J. Neurosci.* **31**, 17488–17495 (2011).
55. Forstmann, B. U. et al. Striatum and pre-SMA facilitate decision-making under time pressure. *Proc. Natl Acad. Sci. USA* **105**, 17538–17542 (2008).
56. Röhner, J. & Ewers, T. How to analyze (faked) Implicit Association Test data by applying diffusion model analyses with the fast-dm software: a companion to Röhner & Ewers (2016). *Quant. Methods Psychol.* **12**, 220–231 (2016).
57. Schmidt, K., Stevens, A., Szabelska, A., Graham, J., Hawkins, C. B., & Nosek, B. The Ideology 2.0 Dataset. *OSF* https://osf.io/2483h (2022).
58. Röhner, J., Thoss, P. & Schütz, A. Lying on the dissection table: anatomizing faked responses. *Behav. Res.* **54**, 2878–2904 (2022).
59. Furr, R. M. & Bacharach, V. R. *Psychometrics: An Introduction* (SAGE, 2014).
60. Nier, J. A. How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Process. Intergroup Relat.* **8**, 39–52 (2005).
61. Nosek, B. A., Greenwald, A. G. & Banaji, M. R. Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Pers. Soc. Psychol. Bull.* **31**, 166–180 (2005).
62. Klauer, K. C., Schmitz, F., Teige-Mocigemba, S. & Voss, A. Understanding the role of executive control in the Implicit Association Test: why flexible people have small IAT effects. *Q. J. Exp. Psychol.* **63**, 595–619 (2010).
63. Grange, J. A. & Schuch, S. A spurious correlation between difference scores in evidence-accumulation model parameters. *Behav. Res.* **55**, 3348–3369 (2022).
64. Krajbich, I. Decomposing implicit bias. *Psychol. Inq.* **33**, 181–184 (2022).
65. Voss, A., Nagler, M. & Lerche, V. Diffusion models in experimental psychology: a practical introduction. *Exp. Psychol.* **60**, 385–402 (2013).

66. Wald, A. *Sequential Analysis* (Wiley, 1947).
67. Carpenter, B. et al. Stan: a probabilistic programming language. *J. Stat. Soft.* **76**, 1–32 (2017).
68. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).
69. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Routledge Academic, 1988).
70. Kruschke, J. K. & Liddell, T. M. The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon. Bull. Rev.* **25**, 178–206 (2018).
71. Morey, R. D. & Rouder, J. N. Bayes factor approaches for testing interval null hypotheses. *Psychol. Methods* **16**, 406–419 (2011).
72. LaFollette, K., Rubez, D., Demaree, H. & Goldenberg, A. Challenging the mechanism for the implicit association test. *OSF* https://doi.org/10.17605/OSF.IO/E97RF (2023).

Acknowledgements

The authors received no specific funding for this work. We acknowledge the data collected by Project Implicit and the Ideology 2.0 team, as well as K. Schmidt for their correspondence. We note here that one statement was amended and one statement was added in the introduction for accuracy at Stage 2, following reviewer feedback: "However, these attempts mainly focused on examining either changes in ease and caution between blocks, or the association between decision ease and the *D*-score, but not response caution as a separate predictor of the *D*-score", which originally read, "However, these attempts mainly focused on examining the association between decision ease and the D-score without paying enough attention to response caution as a separate predictor of the D-score." We added, "Decision ease may also be affected by temporary response strategies that mimic shifts in mental associations, such as faking[30,56]. However, these effects probably stem from task-specific adaptations rather than genuine changes in underlying associations. Thus, decision ease may serve as a more reliable indicator of unconscious processes under typical conditions, while response caution appears at least partially under conscious control."

Author contributions

Study conception and design: K.J.L., D.R., H.A.D. and A.G. Analysis and interpretation of results: K.J.L. Draft manuscript preparation: K.J.L., D.R., H.A.D. and A.G. All authors reviewed the results and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-026-02439-y.

Correspondence and requests for materials should be addressed to Kyle J. LaFollette.

Peer review information *Nature Human Behaviour* thanks Jessica Röhner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Kyle LaFollette

Last updated by author(s): 2/26/2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	N/A
Data analysis	Model fitting was conducted using Stan version 2.25. Analyses were conducted using the following Python libraries: pandas (v2.2.2), numpy (v1.26.4), scipy (1.16.2), matplotlib (v3.10.6), and statsmodels (v0.14.5). Visualizations were conducted using seaborn (v0.13.2) and arviz (0.20.0). All code is available on the Open Science Framework at https://osf.io/2r3zd/, under the Github tab. This can also be found directly at https://github.com/GoldenbergLab/model-iat-drm-test-mechanisms-kyle.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data and materials are available on the Open Science Framework at https://osf.io/2r3zd/.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Participants reported gender in the Ideology 2.0 dataset. However, there were no a-priori hypotheses or theoretical motivation to assume gender differences, so no analyses using gender were preregistered at Stage 1 and thus are not reported. Gender breakdowns by sample are included in Table 3.
Reporting on race, ethnicity, or other socially relevant groupings	Self-reported race and ethnicity were collected as part of the Ideology 2.0 dataset. However, there were no a-priori hypotheses or theoretical motivation to assume race or ethnicity differences, so no analyses using race or ethnicity were preregistered at Stage 1 and thus are not reported. Race and ethnicity breakdowns by sample are included in Table 3.
Population characteristics	No covariate-relevant population characteristics.
Recruitment	Our analyses use data collected from the Ideology 2.0 Study on Project Implicit from December 2007 to June 2012. Data collection was approved by the University of Virginia's Institutional Review Board (IRB) for Social and Behavioral Sciences and informed consent was obtained from all participants. Our report pertains to a secondary analyses and the authors were not involved in data collection nor recruitment.
Ethics oversight	Data collection was approved by the University of Virginia's Institutional Review Board (IRB) for Social and Behavioral Sciences and informed consent was obtained from all participants. The Case Western Reserve University IRB determined that further IRB approval was not required to conduct secondary analyses on these data.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This was a quantitative study (secondary analysis of archival behavioral data), using trial-level response times/choices and explicit preference ratings analyzed with computational modeling and regression-based inferential statistics.
Research sample	Data came from the Project Implicit 2.0 Study (2007-2012) and included 115,601 unique IAT sessions spanning 39 IAT topics (each topic: >= 2,673 participants, mean = 2,964). Participation occurred online via the Project Implicit platform and reflects a self-selected convenience sample of individuals who chose to complete an IAT; thus it is not intended to be population-representative. Because the study was administered remotely via the Project Implicit platform, no researcher was present during task completion beyond the participant. Assignment to study design (Design A vs. Design B) and, where applicable, selection of IAT topic(s) were implemented by the platform using random assignment procedures described in the original dataset documentation. The present work is a secondary analysis of de-identified archival data; thus, investigator blinding during data collection is not applicable, and analyses were conducted according to preregistered hypotheses using a held-out confirmatory sample. Demographic information was self-reported in the archival dataset (when available) and is reported in Table 3."
Sampling strategy	This was a convenience/self-selected sample collected by Project Implicit. Our analyses used a preregistered confirmatory held-out subsample that was masked at Stage 1, with expected effect sizes informed by a separate exploratory/pilot subsample. No new sampling was conducted for the present secondary analysis.

Data collection	Data were collected online by Project Implicit using standard IAT procedures. Data collection was fully computerized/remote with no in-person experimenter.
Timing	Data collection occurred December 2007 through June 2012 as part of the Project Implicit Ideology 2.0 study.
Data exclusions	All exclusions are disclosed in the main text. Exclusions were specified in the preregistered analysis plan and applied during preprocessing and model fitting. Error trials were excluded from RDM fitting because only the corrected (final) RT was recorded, confounding error RTs with multiple decision processes. Trials with RT < 200 ms (likely false starts) and RT > 5000 ms (likely distraction) were excluded. Participants with no within-block response variability (e.g., exclusively pressing one key within any block) were excluded because RDM estimation requires variability to identify drift parameters. We did not count these as completed sessions in the overall dataset, and therefore did not consider them as part of the N = 115,601 confirmatory sample. That said, there were 24,743 such cases in the raw dataset that would have been viable otherwise. No participant or trial exclusions were made via outlier-detection rules beyond the RT bounds specified above. Additionally, subject-level parameters used in explicit-preference regressions were estimated only for participants who had both IAT data passing inclusion criteria and an explicit preference measure for that IAT topic. A total N = 37,023 lacked accompanying explicit preference data in our sample, leaving us with N = 78,578 participants for subject-level RDM parameter estimates." This final sample reflects the sum of IAT samples that has been included in the "Sample with IAT & Explicit Preference" column of Table 2 since Stage 1 acceptance.
Non-participation	Because this is a secondary analysis of an archival online dataset, dropout counts and reasons are not available to the authors and thus are not provided in the manuscript materials. Analyses reflect participants who completed an IAT session with data meeting inclusion criteria.
Randomization	Randomization occurred within the original Project Implicit study procedures. Participants were randomly assigned to Design A vs. Design B. In Design B, participants were randomly assigned to complete two IATs, with topics selected from a pool.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A