Behavior Research Methods
https://doi.org/10.3758/s13428-023-02141-1

Improving the reliability and validity of the IAT with a dynamic model driven by similarity

Peter D. Kvam[1] · **Louis H. Irving**[1] · **Konstantina Sokratous**[1] · **Colin Tucker Smith**[1]

Accepted: 2 May 2023

Abstract
The Implicit Association Test (IAT), like many behavioral measures, seeks to quantify meaningful individual differences in cognitive processes that are difficult to assess with approaches like self-reports. However, much like other behavioral measures, many IATs appear to show low test-retest reliability and typical scoring methods fail to quantify all of the decision-making processes that generate the overt task performance. Here, we develop a new modeling approach for IATs based on the geometric similarity representation (GSR) model. This model leverages both response times and accuracy on IATs to make inferences about representational similarity between the stimuli and categories. The model disentangles processes related to response caution, stimulus encoding, similarities between concepts and categories, and response processes unrelated to the choice itself. This approach to analyzing IAT data illustrates that the unreliability in IATs is almost entirely attributable to the methods used to analyze data from the task: GSR model parameters show test-retest reliability around .80-.90, on par with reliable self-report measures. Furthermore, we demonstrate how model parameters result in greater validity compared to the IAT *D*-score, Quad model, and simple diffusion model contrasts, predicting outcomes related to intergroup contact and motivation. Finally, we present a simple point-and-click software tool for fitting the model, which uses a pre-trained neural network to estimate best-fit parameters of the GSR model. This approach allows easy and instantaneous fitting of IAT data with minimal demands on coding or technical expertise on the part of the user, making the new model accessible and effective.

Keywords Implicit attitudes · Validity · Conceptual similarity · Individual differences · Retest-reliability

Improving the reliability and validity of the IAT with a dynamic model driven by similarity

For more than two decades, researchers have used the Implicit Association Test (IAT) to measure psychological constructs in a way that circumvents the need for introspection on the part of respondents (Greenwald et al., 1998). Most commonly, researchers use IATs in an attempt to capture so-called "implicit" constructs that are generally conceived of as evaluations or beliefs that are relatively uncontrollable and whose existence or influence operates at least partly outside of conscious awareness [e.g.,][] (De Houwer, 2006). The IAT paradigm has undoubtedly been influential outside of academia, with more than 28 million IATs completed at the Project Implicit website (Ratliff and Smith, 2021). The psychometric value of behavioral measures, however, stems largely from their ability to reliably assess meaningful psychological constructs. Critics have long argued the IAT does not succeed in either of those requirements [e.g.,] (Fiedler et al., 2006; Schimmack, 2021; Blanton et al., 2006).

For readers unfamiliar with IATs, the task is structured as follows. Participants are presented with stimuli from two conceptual and two attribute categories – which may include words, pictures, phrases, or other visual or lexical items – and asked to match them with their corresponding category by pressing one of two keys on the keyboard (e.g., 'e' or 'i' for 'left' and 'right' category responses). The task proceeds in a set of four types of blocks, each of which has different sorting rules. Participants begin with two practice blocks, sorting positive and negative words (valence stimuli) and then words related to the categories of interest (e.g., faces of young and old people). In these blocks, participants simply have to classify each stimulus by pressing one button or the other (e.g., positive words on the left, negative words on the

✉ Peter D. Kvam
pkvam@ufl.edu

[1] Department of Psychology, University of Florida, Florida, USA

Published online: 05 July 2023

right). The key manipulation of the IAT comes in the remaining two blocks (referred to as critical blocks) in which the two practice blocks are combined so that there are two categories on the left and two categories on the right side of the screen. In these blocks, participants use a single response key to sort evaluative stimuli (e.g., positive words) and category stimuli (e.g., faces of young people). The task is illustrated on the left side of Fig. 1. The guiding idea behind the IAT is that responses will be easier (i.e., faster) when categories that share a response key also share a relationship in the participant's mind and more difficult (i.e., slower) when categories sharing a response key do not share a relationship or are even at odds with one another. In other words, researchers use an individual's pattern of responses on the IAT to draw conclusions about, for example, their degree of positivity toward one social group relative to another.

Of course, there are research practices such as selecting appropriate stimuli and using multiple measurement occasions that can improve on an IAT's psychometric properties (Greenwald et al., 2021; Carpenter et al., 2022), but there are certainly serious and credible concerns about the structure of the task and its ability to measure individual differences. In aggregate (e.g., averaging scores by experimental condition or geographical region), many IATs reliably produce large effects, distinguish known-groups, and are associated with relevant outcomes (Payne et al., 2017). Where the IAT could stand to improve the most is at the individual-level. The existing research indicates low test-retest reliability (Gawronski et al., 2017) along with imprecise individual estimates (Klein, 2020), reflecting psychometric properties that make it potentially undesirable as a measure of individual differences. Moreover, the predictive validity of IATs is also disputed. On one hand, some researchers interpret the available evidence as indicating IATs can predict an array of relevant outcome measures (Buttrick et al., 2020; Greenwald et al., 2009; Kurdi et al., 2019), in some cases over and above analogous self-report measures. On the other hand, critics argue that there is little to no evidence that IATs meaningfully predict any outcomes with practical relevance or real-world significance at all (Carlsson and Agerström, 2016; Van Dessel et al., 2020), and not incrementally over any sufficiently valid self-report measure (Blanton et al., 2016; Oswald et al., 2013). We want to continue to remind readers that the IAT is a measurement procedure in much the same way that a survey is a measurement procedure and that there is no such thing as The IAT. It may be, for example, that psychometric properties are better for IATs measuring some attitude objects rather than others.

Designing effective behavioral measures is difficult in part because it requires balancing a trade-off between the robustness of experimental manipulations (i.e., how consistently a manipulation creates an effect) and the reliability for assessing individual differences in task performance [suggestthatotherdimensionsormeasuresofreliabilitymaynotbeasconcerningfortheIAT] (Greenwald et al., 2021). This issue is not unique to IATs, and has been termed the "reliability paradox" (Hedge et al., 2018). However, Haines

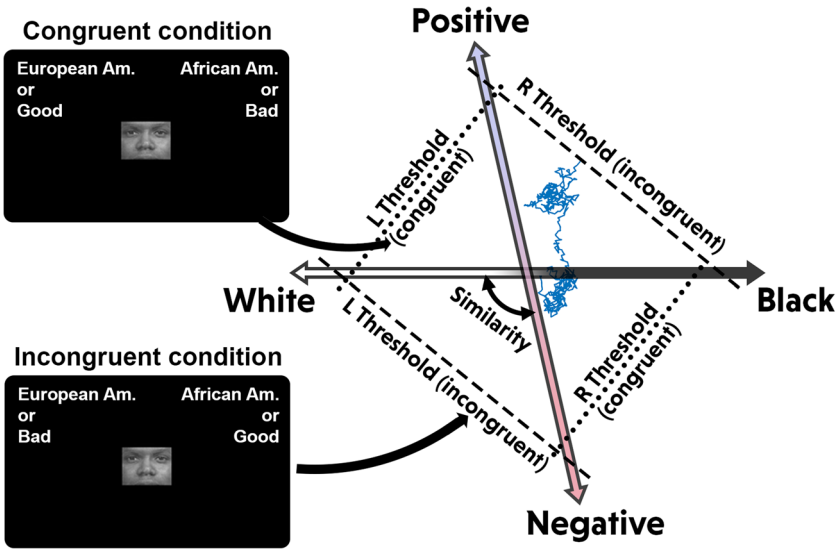

Fig. 1 Diagram of the structure of the GSR-DDM model. The stimulus provides information that guides participants toward white / black and positive / negative evaluations, which are mapped onto responses on the left (L) or right (R) sides of the screen based on which threshold (dotted / dashed lines for compatible / incompatible blocks, respectively) is crossed. This drives an evidence accumulation process, shown in blue, that moves around until it hits one of the choice boundaries. Shown here is a model of an individual with a slight bias toward white faces (smaller angle between white faces & positive and black faces & negative) relative to black faces (larger angle between black faces & positive and white faces & negative)

et al. (2021) have shown that this issue is one of *measurement* as opposed to an issue with behavior in general. Specifically, they showed that the way behavior is quantified on tasks like IATs is counterproductive to high reliability and, consequently, predictive validity. There are two reasons for this. First, simple summary statistics like mean response times do not fully capture the rich patterns of behavior that people exhibit on these tasks, lacking the distributional information about response times as well as the accompanying accuracy participants exhibit. By condensing performance down to a single index like the *D*-score (Greenwald et al., 1998, 2003), all the different cognitive processes involved in performance – and the error in measurement – are confounded.

As we note below, there have been efforts to dissect performance into multiple dimensions with models like the diffusion model (Röhner and Lai, 2021; Klauer et al., 2007), but even these models suffer from an additional issue related to test-retest reliability. Specifically, they quantify performance on IATs using a difference score or comparison between conditions. Any time performance is quantified in two separate conditions and then compared between them, the summary statistic or model parameter used to quantify behavior in any single condition is doubled. This issue has been raised before in the psychology and educational measurement literature (Bereiter, 1963; Thomas and Zumbo, 2012; Overall and Woodward, 1975; Gardner and Neufeld, 1987), but it was dismissed or glossed over largely because tests of significance between conditions (e.g., blocks of the IAT) based on the difference score are still well-powered [and in fact they are *highest* when reliability is zero] (Overall and Woodward, 1975). This means that comparisons between conditions are likely to yield significant results, but will not serve well as reliable measures of individual differences.

A natural parallel can be drawn to IATs, where large effects of pairing manipulations (attitude-congruent / attitude-incongruent) are observed alongside low test-retest reliability (Gawronski et al., 2017). This trade-off is potentially damaging to IATs because its theoretical underpinnings are largely predicated on its ability to measure differences in individual-level automatic cognitive processes (Greenwald et al., 1998; Kurdi and Banaji, 2017). Its value as a behavioral measure is in its ability to assess these latent processes, making reliable measurement and high predictive validity paramount to effective use. To the extent that the test-retest reliability and predictive validity of IATs is diminished by measurement practices, we should strive to improve our measurement procedures to imbue the task with greater utility. To use it as a measure of individual differences, we must therefore solve the dual challenges of generative modeling and avoiding relying on difference scores. This paper outlines a modeling approach that accomplishes both objectives.

Modeling IATs

One of the reasons that IATs have been criticized is that the usual metric for summarizing IAT task performance (IAT *D*-scores) is inappropriately interpreted as a process-pure measure of individual differences in automatic associations (Conrey et al., 2005; Schimmack, 2021). It is fairly clear that, although IATs almost certainly pick up on associative relationships, a purely associative account is not tenable and, instead, the possibility exists that IATs also pick up on propositional information [for the most recent and comprehensive account see] (De Houwer et al., 2021). In other words, researchers assume *D*-scores primarily capture the strength of target-attribute associations stored in long-term memory rather than ephemeral or non-associative factors see (Fiedler et al., 2006; Bading et al., 2020). However, behavior on an IAT does not correspond one-to-one with the (automatic) activation of underlying attitudes (e.g., associations). Rather, it results from a mix of controlled and automatic processes see (Calanchini and Sherman, 2013) and contains both attitudinal and non-attitudinal content (Calanchini et al., 2014). For example, IAT task performance is influenced systematically by non-associative cognitive variables including general processing speed, task-switching abilities, and cognitive control [e.g.,] (Blanton et al., 2006; Ito et al., 2015; Klauer et al., 2010).[1] Additionally, different IATs that target seemingly distinct attitudes still have substantial overlap in their associations after decomposing IAT *D*-scores into associative and non-associative components (Calanchini et al., 2014), thus indicating the presence of construct-irrelevant, common method variance or attitudinal content that is irrelevant to the specific constructs of interest.

We posit that the controversies surrounding the reliability and validity of IATs are intractable until researchers embrace modeling approaches that can decompose the individual-level behaviors into unique components that are both reliably quantifiable and theoretically-grounded [e.g.,] (Klauer et al., 2007). Doing so will allow researchers to gain new insight into which specific aspects of IAT task performance can be reliably captured across repeated measurements (i.e., test-retest reliability), and whether the unique parameters can predict outcomes above and beyond typical *D*-scores (i.e., predictive validity). The complexity underlying IAT task performance ought to be accompanied by scoring metrics that can meaningfully capture and distinguish the multiple underlying processes. Unfortunately, the simple metrics and models traditionally used to characterize behavior [such as

[1] Of note, reducing the impact of these cognitive variables was one of the central improvements of the IAT *D*-score over the original scoring procedure which consisted of unstandardized differences between block means (Greenwald et al., 2003; Cai et al., 2004; Mierke and Klauer, 2003)

the IAT *D*-scores, *C*-scores, or other summary statistics; for an overview, see] (Röhner and Thoss, 2019) provide an impoverished view.[2]

Recent evidence suggests that the IAT's utility as a trait-level measure can be increased greatly by requiring individuals to complete at least two IATs and aggregating *D*-scores onto a latent variable (Carpenter et al., 2022). This approach has real promise, but retains the *D*-score, which we argue is not ideal. Indeed, the field's near total reliance on IAT summary statistics and scores has left researchers to make inferences about "implicit" constructs that are, arguably, too ill-defined and heterogeneous – not only between different researchers but also within the same ones – to have real theoretical and practical utility (Corneille and Hütter, 2020; De Houwer et al., 2009; Gawronski et al., 2022; Schimmack, 2021). Furthermore, the repetition of IATs as in the procedure described in Carpenter et al. (2022) increases the risk of participants learning to fake their performance in an attempt to mask their attitudes from detection via traditional scoring methods (Röhner et al., 2011; Fiedler and Bluemke, 2005). As we note below, models are better able to disentangle faking strategies from activated associations, presenting an additional solution to the repetition problem.

In this paper, we address this issue by developing a new computational model of performance on IATs, adopting and formalizing the link between models of semantic meaning and similarity and models of decision-making on response time tasks. The goal of our modeling approach is to disentangle the many factors influencing performance on IATs – including response caution, encoding, and non-decision processes such as pushing the keys on the keyboard (Ratcliff et al., 2016; Busemeyer et al., 2019) – from the construct-relevant mental content that the task is designed to measure. To accomplish this, we estimate the cosine-similarity between the concepts specific to an IAT (e.g., Black, White, good, and bad for the Black/White Race IAT) alongside model parameters describing other elements of the decision process. A complete detailing of this model is provided in the "Modeling approach" section, but we provide a summary of the benefits here. Specifically, our modeling approach provides a richer description of what participants *are actually doing* on an IAT by providing multiple measures quantifying performance on the task. Applying a cognitive modeling perspective to IATs provides a relatively clean theoretical slate. In particular, attempting to define the "implicit" nature of the associations captured by the task a priori (Corneille and Hütter, 2020; Gawronski et al., 2022) is unnecessary to the approach. Measuring representational similarity in terms of geometric relationships makes no assumptions about how similarities are formed or whether participants are aware of their existence. We merely presuppose that these semantic similarities exist and affect decisions, an assumption that has been well-validated [e.g.,][] (Bhatia, 2013, 2017; Bhatia and Mullett, 2018) to the extent that it should be reasonable even to the strongest critics of IATs.

At their core, IATs are designed to measure the strength of relationships (e.g., associations) between categories (e.g., social groups) and attributes such as valence (Greenwald et al., 2021) or personality traits (Back et al., 2009). In this respect, they are similar to vector space semantic models, which seek to represent the meanings of words and concepts in terms of the similarities between them (Landauer and Dumais, 1997; Günther et al., 2019), including the representation of collective biases reflected in language (Charlesworth et al., 2021). In typical semantic models, these similarities are measured from word co-occurrences across large samples of text (Turney and Pantel, 2010) ranging from 100,000 documents to 1-2 billion for large language models like BERT (Zhu et al., 2015) or even half a trillion tokens for models like GPT-3 (Brown et al., 2020). One of the architects of the IAT, Greenwald (2017), noted the potential relationship between IATs and vector-space models seeking to represent similarity between concepts: "Caliskan et al. (2017)'s Word-Embedding Association Test (WEAT) algorithm uses cosine similarity (a correlation-like indicator) between word vectors in different word categories, much as the IAT uses response latencies; greater cosine similarity corresponds to faster IAT responding."

The idea that response times on IATs reflect semantic similarities among words, or between words and visual stimuli is intuitively appealing. The speed at which we can retrieve the word "royal" from the word "gold" is much faster than we can retrieve it from the word "shale" – which might seem unassociated or even inversely associated with royalty. When these associations are pitted against one another, it can create competing or interfering relationships among categories. For example, if the categories were royal/gems vs peasant/rocks, we might expect a "ruby" stimulus to easily correspond to the former, resulting in a fast response; however, when they are juxtaposed or conflicting as in royal/rocks vs peasant/gems, the royal and gems categories might compete to make the response to "ruby" slower. Low conflict stimuli, trials, and conditions should create less interference and thus better (faster) performance, whereas conflicting or incongruent semantic similarities should result in worse (slower) performance. In this way, the degree of competition or interference is thought to provide the link between semantic associations on one hand and response speed on the other.

[2] It is worth pointing out that there have been positive aspects to having the IAT *D*-score serve as a field-standard metric, particularly in comparison to tasks such as evaluative priming in which there are many different scoring procedures. For example, it is relatively intuitive, and has facilitated ease of comparison across disparate data collections.

However, IAT performance involves many processes beyond similarity: a decision-maker performing the task must encode the stimuli (a process which may be faster or slower depending on the stimulus type, such as word vs face), relate the stimuli to support for the different response categories, determine what response to trigger, and carry out the action of entering their response. The modeling challenge is to disentangle these processes from those related to associations between concepts, determining if and how similarity has impeded or facilitated the choices that someone makes across trials and conditions of IATs.

Critics and proponents of IATs agree that better modeling techniques are the most promising path forward to developing an accurate account of the multiple cognitive processes that generate task performance (Carpenter et al., 2022; Schimmack, 2021; Corneille and Hütter, 2020; Gawronski, 2019; Fiedler et al., 2006). Applying cognitive models to behavioral tasks allows for effective estimation of distinct latent cognitive processes underlying performance on response time tasks in social cognition (Pleskac et al., 2018; Johnson et al., 2017). Thus, it becomes possible to reliably characterize individual differences in latent processes by developing quantitative theories of how conceptual similarities influence choices and response times.

Indeed, researchers have already begun applying various approaches to generative models of the behaviors underlying IAT task performance. In the next section, we outline how several of these approaches have improved measurement practices on IATs, then segue into our own approach and how it solves many outstanding issues with IAT modeling.

Multinomial process trees

Two of the earliest and most common models for decomposing IAT task performance, the quad (Wang et al., 2019; Dunham et al., 2016; Ruiz et al., 2015; Wrzus et al., 2017) and ReAL (Meissner and Rothermund, 2015; Jin, 2016; Koranyi and Meissner, 2015; Calanchini et al., 2021) models, use multinomial process trees [MPTs] to account for accuracy data on IATs. In these models, error rates are compared across conditions of the task to make inferences about the order in which different cognitive processes occur (Hütter and Klauer, 2016). The first and most well-known model of IAT behavior is the Quad model [i.e., Quadruple Process model:] Conrey et al. (2005) which uses the distribution of error responses to estimate four different parameters. Namely, representing activation of target-attribute associations (estimated separately for each IAT block), accuracy in detecting correct responses, self-regulation to overcome associations that would result in incorrect responses, and guessing when other processes fail to fully guide responding (Calanchini and Sherman, 2013). The activation parameter thus represents relatively process-pure associations whereas the remaining parameters represent non-associative or mixed processes. The model allows researchers to determine the extent to which each process guides IAT task performance to answer various research questions.

A second model, the ReAL model, estimates three parameters that represent a task-simplifying recoding process, activation of evaluative associations (separately per IAT block), and label-based discrimination of the correct response (Meissner and Rothermund, 2013). The recoding parameter is central to the model because it estimates the role of a wide variety of response strategies that are distinct from evaluative associations that are typically assumed to underlie IAT effects. In general, participants are more likely to apply recoding to the compatible[3] IAT block by reducing the target and attribute categories into a single category. Thus, recoding is problematic for typical interpretations because it is a non-associative process that causes IAT effects to appear more stereotype-consistent (e.g., stronger preference for majority over minoritized social groups).

Both the Quad [e.g.,] Wang et al. (2019); Dunham et al. (2016); Ruiz et al. (2015); Wrzus et al. (2017) and ReAL model [e.g.,] Meissner and Rothermund (2015); Jin (2016); Koranyi and Meissner (2015); Calanchini et al. (2021) have shed light onto the cognitive processes underlying performance on IATs, and generated insights that would not have been possible without these modeling approaches. For example, research using the ReAL model suggests that recoding is responsible for producing the smaller IAT effects that are observed with word- versus picture-based stimuli (Meissner and Rothermund, 2015) as well as the apparent differences in gender associations for younger versus older males (Jin, 2016). After controlling for recoding, the ReAL model estimates the unique contributions that positive and negative associations between target concepts and attributes make to IAT scores. Researchers can then not only be more confident that they are making inferences about associations per se, but also about the specific content of those associations. For example, the ReAL model provided evidence that motivation to initiate romantic relationships leads specifically to weaker associations between potential partners and negative characteristics (Koranyi and Meissner, 2015). Similarly, the Quad model illustrated that performance on the Young-Old IAT differs by gender, race, and motivation to control prejudice (Ruiz et al., 2015), and age-related effects across

[3] Note that we use the term "compatible" or "congruent" to describe conditions where a minoritized group is paired with a negative valence category while a majority group is paired with a positive one, and "incompatible" or "incongruent" to refer to conditions where the minoritized group / stimuli are paired with positive-valence words. In other words, it is "compatible" with dominant cultural attitudes at the time of writing, but does not indicate that it is "compatible" with truth or with the attitude of any individual participant.

various IATs appear to be attributable primarily to differences in overcoming bias rather than in association strength (Wrzus et al., 2017). The ReAL model is used infrequently despite its potential applications, perhaps because it requires several modifications to the typical IAT procedure including an increased number of trials. For example, Calanchini et al. (2021) applied both the original ReAL model and a simplified version across a wide set of IAT procedures (e.g., 320 vs. 120 trials, single block vs. multi-block). The original ReAL model performed well across a number of conditions but cannot be fit to the standard IAT procedure; the simplified model was also unable to provide a range of meaningful parameter estimates from standard IAT data. However, the Quad model can be fit to standard IAT data, provided there are enough errors. Below, we apply the Quad model and compare it to our new approach in order to better situate it relative to multinomial processing trees.

These models can provide unique insights into IAT task performance in part because typical response time measures like IAT *D*-scores omit response accuracy. Nevertheless, multinomial process trees fail to overcome a major deficiency of IAT *D*-scores because they omit a different source of information (response times) that could shed light on the underlying cognitive processes. Although some different formulations of the *D*-score attempt to integrate the two sources of information into a single metric, such as adding a response time penalty for errors (Röhner and Thoss, 2019), they cannot fully account for both at the same time. As a result, both sum scores and Quad / ReAL models are unable to detect the joint information that is provided at the intersection of response times and accuracy, such as if a participant responds more slowly to improve accuracy or when a participant sacrifices accuracy to respond faster (Luce, 1986; Röhner et al., 2013). People commonly differ in how they approach these so-called speed-accuracy trade-offs when completing response time tasks (Wickelgren, 1977; Heitz, 2014), so models that cannot or do not capture this trade-off are almost certainly missing a fundamental piece of the behavioral phenomenon of completing an IAT.

Diffusion model

A classic and effective approach to modeling binary choice response time tasks, such as those in IATs, is to use a dynamic decision-making model like the diffusion model, where noisy information is accumulated over time until a decision threshold is reached and a response is initiated (Ratcliff, 1978; Ratcliff et al., 2016; Ratcliff and McKoon, 2008). A basic version of the diffusion model (i.e., not including start point bias or cross-trial variability) estimates parameters related to three processes underlying IAT performance, including the speed at which information is gathered (drift), speed-accuracy tradeoffs (thresholds), and nondecision time (Röhner and Lai, 2021; Röhner and Ewers, 2016; Klauer et al., 2007). The accumulation speed parameter or *drift* represents how easily stimulus information is processed before deciding which key to press on each trial, and it is intended to reflect the relatively automatic associative processes between targets and responses that researchers typically hope to measure with IATs. The speed-accuracy tradeoff parameter (threshold) estimates the level of caution people take to ensure accurate rather than quick responding. It is often considered a construct-irrelevant decision-making process for implicit measures, because it is thought of as relatively controlled rather than automatic process, although it has been demonstrated to be related to method-specific variance as well as faking strategies (Klauer et al., 2007; Röhner, 2016) and it is actually possible that thresholds are learned or adjusted over time, which may occur automatically or without awareness following reinforcement (Cavanagh et al., 2011; Fontanesi et al., 2019b). Non-decision time captures variance in IAT performance stemming from individual differences in cognitive and motor abilities, as well as any other external factors that cause people to increase or decrease their response speed. As a result, the diffusion model distinguishes between processes that are directly related to decision making (drift / v, threshold / a) and processes that can be seen as outside the decision process (non-decision time / $t0$) (Röhner, 2016).

A great deal of progress on modeling IATs using the diffusion process has been made, often with the aim of disentangling processes of interest (e.g., automatic activation of evaluative associations) from processes deliberately controlled by a participant (Röhner et al., 2013; Röhner and Ewers, 2016; Röhner and Thoss, 2018; Röhner and Lai, 2021; Röhner et al., 2022; Klauer et al., 2007; van Ravenzwaaij et al., 2011b; von Krause et al., 2021). A strength of this approach is that it is able to disentangle construct-related variance in the drift from impression management and "faking" strategies, which often appear as shifts in thresholds or non-decision times between congruent and incongruent conditions (Röhner and Ewers, 2016; Fiedler and Bluemke, 2005). It has also succeeded in dissecting the processes underlying the impact of interventions on IAT performance (Röhner and Lai, 2021), and the effects of modifying IAT target categories and stimuli (van Ravenzwaaij et al., 2011b).

The limitation of traditional diffusion modeling methods is that the parameters are estimated separately for congruent and incongruent conditions – and often neglect the other conditions entirely. Once computed for each condition, a difference score such as IAT_v (difference in drift rates between conditions), IAT_a (difference in thresholds between conditions), or IAT_{t0} (difference in non-decision time between conditions) is computed. As we mention above, these difference scores are effective for detecting differences in behavior between conditions, but they will naturally be unreliable

because of the compound error variance (Bereiter, 1963; Thomas and Zumbo, 2012; Overall and Woodward, 1975; Gardner and Neufeld, 1987).

As with the ReAL model, diffusion model analyses are typically too complex to be applied with less than 90 trials per condition (Röhner and Ewers, 2016; Klauer et al., 2007). One way to overcome this limitation is by using the Discrimination-Association Model to estimate similar parameters but with a mathematically simpler Poisson race model (Stefanutti et al., 2013) or to use a simplified version of the diffusion model like the E-Z diffusion model (Paige et al., 2022; Wagenmakers et al., 2007; Röhner and Thoss, 2018). These approaches are advantageous over the diffusion model in that they need less information and can function even when there are no trials with incorrect responses, but non-identifiable parameters remain common when analyzing standard IAT data. However, all of them still require a reasonably high level of coding and modeling ability to apply. Although computational modeling ought to be an accessible and achievable route to better theory in psychology (Guest and Martin, 2021), it is often avoided because of the demands on quantitative and programming skills on the part of the model user. We suspect that this barrier has significantly affected research on IATs, where many may wish to use computational models but lack the background to confidently do so.

Technical issues with model estimation aside, each of the variants of the diffusion model suffer from a major disadvantage in that they do not directly index what IAT researchers are usually interested in – capturing the degree of similarity or association between concepts (e.g., Black / White faces and positive / negative words). The models are applied such that they estimate separate drift rates for the compatible and incompatible IAT blocks, meaning the only way to get a proxy for degree of similarity is by contrasting the parameter estimates. As we mentioned above, this falls prey to the second issue of reduced reliability in behavior, which is the compound error of difference scores. This issue is exacerbated by the fact that the drift rate measures processes above and beyond the activation of associations, such as the ease with which a particular type of stimulus is processed (e.g., words vs faces), the discriminability of the categories, and the relative strength of category activations (Kvam and Pleskac, 2016). The more processes that need to be quantified using the same parameter, such as drift, the less specific – and arguably, less informative – that parameter tends to be. Using the difference between catch-all drift rate parameters therefore results in an estimate that contains greater uncertainty and fails to directly quantify the association-specific processes that IATs are designed to measure. In our analyses in this paper, we show that this results in greatly reduced reliability and ultimately predictive validity.

Despite its drawbacks, the diffusion model is an excellent starting point for building a model of IATs because it quantifies behavior – including both response times and accuracy – in terms of meaningful cognitive processes. The proposed work improves on the diffusion modeling approach by directly addressing the practical and theoretical hurdles outlined above. It specifically addresses the main drawbacks of current modeling approaches, which are

1. Separately computed metrics of performance for congruent and incongruent conditions, resulting in difference scores [e.g., IAT_v, IAT_a, and IAT_{t0};] Röhner and Lai (2021);
2. The absence of a single parameter that directly quantifies conceptual similarities between stimuli and categories;
3. Focusing on response times alone (most *D*-score measures), accuracy alone (Quad, ReAL), or only the congruent and incongruent conditions (Quad, ReAL, diffusion, and *D*-score measures) while ignoring useful information contained in the remaining IAT data;
4. The difficulty of applying complex dynamic cognitive models to the relatively small number of trials typically observed in IAT studies; and
5. The difficulty of applying cognitive models in general.

Our approach addresses each of these issues by (1) quantifying associations between concepts (Black / White, positive / negative) in terms of a single parameter that quantifies individual differences in performance; (2) doing so using a representational similarity framework that predicts the differences in both choice and response time between conditions; (3) leveraging data from all four conditions in order to disentangle differences in performance related to different stimuli versus differences related to experimental manipulations; (4) using hierarchical Bayesian methods for model fitting that help constrain estimates of individual-level performance on IATs even with small sample sizes; and (5) introducing a new online tool for automatically fitting the model we developed to IAT data.

Modeling approach

Our approach to modeling behavior on IATs uses a framework called the geometric similarity representation (Kvam, 2019a; Kvam and Turner, 2021), which generalizes the diffusion decision model to an arbitrarily large number of interrelated choice options. The GSR has been used to model multi-alternative choice and continuous-response paradigms (Kvam, 2019a, b; Kvam and Busemeyer, 2020; Kvam et al., 2023), where it has been subjected to stringent tests like selective influence. Because IATs only feature two response

options (left and right), this generalization of the diffusion model is largely conceptual, as drift rates cannot be disentangled into direction and magnitude when there are only two response options. Fortunately, the implementation of the GSR in binary choice is still conceptually richer as well as computationally convenient, sharing many advantages of the diffusion model while introducing a measure of representational similarity. The similarity metric is derived from computational linguistic models (Deerwester et al., 1990; Landauer, 2006; Landauer and Dumais, 1997; Furnas et al., 1988; Goldberg and Levy, 2014; Lin et al., 1998), where different concepts are represented as vectors in a multidimensional space and the similarity between concepts is a function of the angle between vectors. In doing so, it forges a formal link between these two approaches that has been suggested by others (Greenwald, 2017). We refer to this model as the geometric similarity representation extension of the diffusion decision model, GSR-DDM or simply GSR, reflecting its roots in traditional evidence accumulation models like the DDM while emphasizing its new connections to models of conceptual similarity using vector-space semantics.

The idea underlying the GSR-DDM model is illustrated in Fig. 1: a small angle between two concepts ($<$ 90 degrees), such as White faces and positive, indicates that two concepts are more similar. Conversely, a large angle between concepts ($\geq$ 90 degrees), such as White faces and negative, indicates that they are dissimilar to one another (Kvam, 2019a; Kvam and Turner, 2021; Smith, 2016). The GSR-DDM incorporates these representations of concepts to estimate the angle between (for example) a Black-White faces axis and a negative-positive valence axis as a measure of the relative similarity between the concepts from an IAT.

The similarities among concepts are then built into a model of accuracy and response time using a geometric framework developed by Kvam (2019a). This model predicts the same distributions of accuracy and response times as the diffusion model (a Wiener distribution) but disentangles drift rates into stimulus-specific factors and conceptual similarity among category responses. In this approach, a decision can be facilitated or hindered based on the similarities or associations between the options or features under consideration. As a decision-maker considers their options, their state changes over time according to the attributes or information provided by the stimulus. In the Race IAT, the visual features of an image might favor a "White" or "Black" response, moving their state upward or downward in a space like the one depicted in Fig. 1. The idea motivating IATs as an implicit measure of attitudes is that the image may also carry some positive or negative valence based on a decision-maker's relatively automatic associations with those categories. In GSR-DDM, this is reflected by a smaller angle between concepts of Black (race) and negative (valence) as well as between white and positive. Compatible associations or similarities speed up the decision process by allowing both the target category and its evaluation valence to lead participants toward the same response (e.g., a "left" / L response). Conversely, incompatible associations or dissimilarities slow down the decision process by pulling decision-makers in opposing directions [similar to lateral inhibition in the LCA] ;Usher and McClelland (2001). If the evidence accumulation process crosses a category boundary associated with the target category or its partner (e.g., "Positive" or "White person" for a positive word in the congruent condition), it triggers a correct response. If it instead crosses a category boundary of one of the two other categories (e.g., a "Negative" or "Black person" response for a positive word in the congruent condition), an incorrect response is generated. By connecting these cognitive processes to the frequency of these boundary crossings, both accuracy and response time are used to estimate the parameters of the model.

The decision process for determining which choice is triggered could be described as a mutual power struggle between the responses of Black, White, positive, and negative on a Race IAT. A stimulus showing a positive word tips the balance of power in favor of the positive response, but it may also "activate" or cause the Black or White responses to muster strength according to the conceptual similarity relationships between race and valence. If they associate Black faces with positive valence, then the Black faces and positive valence responses will work in tandem, thus resulting in the fastest response times when Black faces and positive valence are paired on the same side of the screen. Conversely, the Black faces and positive word responses would engage in a direct struggle with one another when Black faces and positive words are on opposing sides of the screen, thus resulting in slower response times.

A confluence of interfering or facilitating activations is a feature present in many psychological tasks. For example, a participant in a Stroop task typically speeds up when the words and colors are aligned relative to a neutral condition where words are unrelated to colors (Heathcote et al., 1991; Lindsay and Jacoby, 1994), indicating that the presence of a facilitating word leads to faster choices. Conversely, colors that conflict with words can be influenced by lexical processing that drives decisions *away* from the correct responses, thus slowing people down in incompatible blocks (MacLeod, 1991, 1992). Like the Stroop task, IATs seek to measure interference or facilitation among stimuli by measuring response times. Using a model of response times that closely resembles the underlying cognitive process allows us to create more reliable and valid measures of association (Haines et al., 2021), enabling the GSR-DDM to substantively improve upon standard practices such as the *D*-score. However, to do so, its assumptions must align with the structure of the neural and cognitive mechanisms that support performance on IATs. There are a few aspects

of the structure of IATs that are relevant for constructing a new model. Specifically, we make three observations about decision-making on a typical IAT:

1. Most IATs feature multiple types of stimuli, such as faces and words, which are processed at different speeds in the brain (Heider and Groner, 1997).
2. Participants can change their decision criteria from condition to condition, either intentionally as part of a deliberate impression management or faking strategy (van Nunspeet et al., 2015; Röhner et al., 2013) or even unintentionally based on feedback or cues to task difficulty (Fontanesi et al., 2019b).
3. Response times for correct responses are typically faster than those for incorrect category responses (which are rare). This is true even when looking at the raw mean response times, or after removing response time penalties for incorrect responses included in some scoring procedures.

The final observation may depend on the specific IAT being used, but appears common from our analyses.[4] These observations inform our modeling assumptions. Based on point (1), we use two separate drift rates for different types of stimuli, in this case words and faces. These two types of stimuli are processed in entirely separate neural circuits of the brain, with face processing occurring in the inferior temporal cortex / fusiform face area (Bentin et al., 1996) and lexical processing of written words occurring in opposite hemisphere (McCandliss et al., 2003). Backward masking studies have made it clear that words and faces are processed at different speeds (Heider and Groner, 1997). Therefore, drift rates that are intended to capture the processing of these two types of stimuli relative to response options should naturally differ according to whether faces or words are being assigned to different categories. As we show later, this distinction in the model is vindicated by substantially higher drift rates for faces than for words across participants and conditions.

The second observation implies that participants can adjust their thresholds strategically, either trading accuracy for speed to reduce response time or taking a longer time on each choice to maintain accuracy (Wickelgren, 1977; Luce, 1986; Heitz, 2014). Accounting for this *speed-accuracy tradeoff* is critical to assessing performance on the IAT, and is one of the main reasons that models like multinomial processing trees (which only assess accuracy) or the *D*-score (which only considers response times) alone cannot provide complete accounts of performance. In GSR-DDM, changes in response caution can occur across conditions. For example, a participant concerned about appearing biased in their response times (Schlenker, 1980; Röhner et al., 2013; Röhner and Ewers, 2016; Röhner et al., 2022) might sacrifice accuracy to match their mean RTs across compatible and incompatible blocks. Allowing thresholds to vary provides the opportunity to involve explicit or conscious processes in performance on an IAT. It also indexes an element of behavior that is orthogonal to similarity (or implicit associations) but still relevant to beliefs and behaviors related to topics an IAT seeks to measure, like race, sexual orientation, age, and so on.

The final observation imposes a restriction on the modeling approach; namely, that it should be able to capture patterns of response times that are asymmetric (faster) for correct relative to incorrect responses. In the modeling approach we adopt, this is accomplished by assuming that the stimuli have random effects (Ratcliff, 1978; Ratcliff and Smith, 2004), i.e., some stimuli yield stronger signals than others. This is a common assumption in signal detection, where the strength of "signal" and "noise" stimuli each follow normal distributions (Green and Swets, 1966). Our measure of signal strength, the drift rate, affects both RT and choice accuracy: higher drift rates result in more correct and faster responses, while lower drift rates result in slower and fewer correct responses. The reason this is able to produce asymmetric response times is that weaker signals (lower drifts) are more likely to result in incorrect responses. As a result, incorrect responses appear to result more frequently from weak or conflicting signals, associated in the model with longer response times, as opposed to fast guesses or strong / variable prior beliefs, which are typically associated with shorter response times.

Put together, the GSR-DDM features (1) a conceptual-similarity parameter that describes the relationships between the concepts on either side of the screen on an IAT; (2) a mean drift rate for each type of stimulus present in the study describing how fast they are processed; (3) a threshold for each condition of the study, controlling how careful a participant is relative to how quickly they wish to decide; and (4) a drift rate variability parameter describing how much variance there is in the drift rates for different stimulus sets (e.g., "positive" stimuli). The effects of each of these parameters on response times and accuracy is shown in Fig. 2. The key element of GSR-DDM that enables it to account for differences in behavior between compatible and incompatible blocks on an IAT is the similarity parameter, which provides a singular measure of how a participant represents the relationship between the concepts activated by stimuli. However, the relative response times and accuracy between conditions are also affected by thresholds (how cautious a participant is) and differences in processing speed between different stimulus sets (drift rates) as well as variability within these stimu-

[4] For example, the Personalized IAT (Olson and Fazio, 2004) does not include error feedback or require the respondent to correct their error responses.

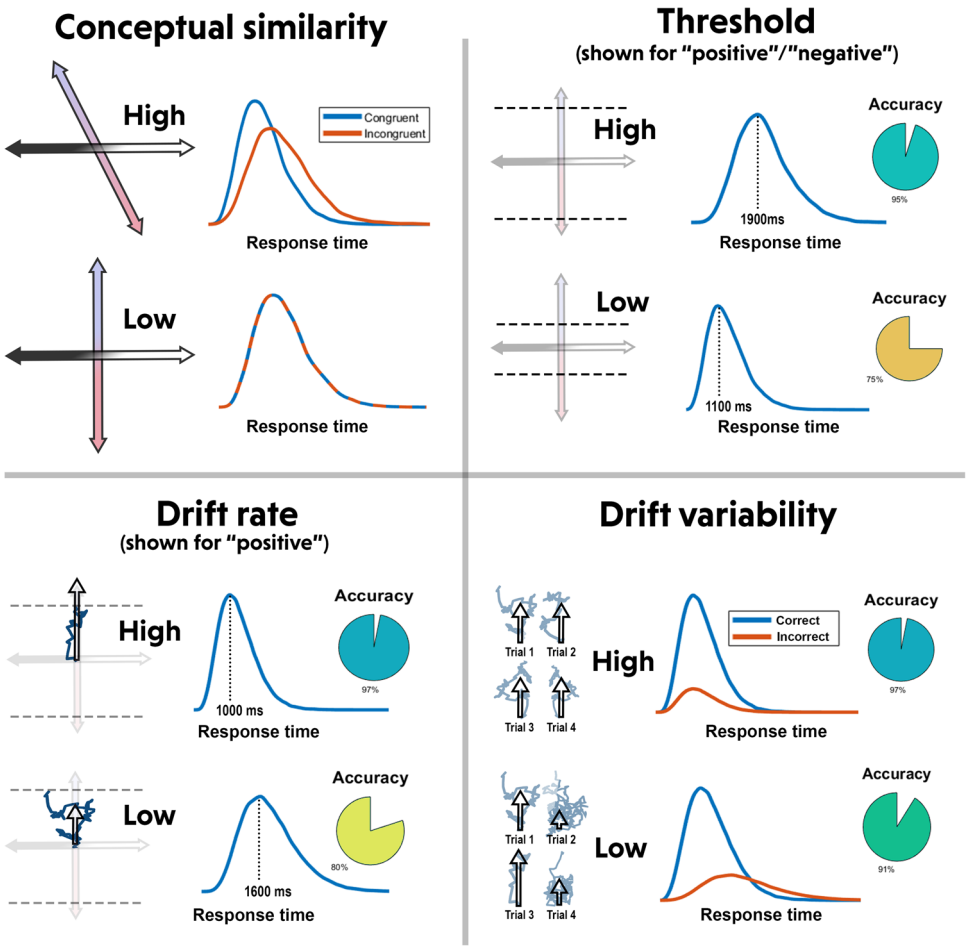

Fig. 2 Effects of manipulating each of the parameters of GSR-DDM (except non-decision time, which simply shifts the response time distribution right or left)

lus sets from trial to trial or stimulus to stimulus (drift rate variability).

One final note is that we are using the relative-evidence choice boundary version of the GSR, where responses are made based on the balance of support between two actions (left, right) as opposed to the absolute degree of support for each of the possible categories (e.g., Black, White, good, and bad) separately. This modeling choice is similar to using a diffusion rather than an accumulator model, although these are only two among many configurations of response boundaries that are possible (Kvam, 2019a; Kvam et al., 2023). Because we only collect a limited amount of accuracy and response time data with each IAT, it would be almost impossible to tell the two approaches apart from one another (Donkin et al., 2011). However, future work looking at confidence judgments (Reynolds et al., 2021), distributions of evidence collected (Kvam et al., 2022), neuroimaging data (Turner et al., 2015), or other information should shed light onto the representations of evidence during decision-making on IATs.

Modeling summary

So far, we have examined a variety of different approaches to modeling performance on IATs and enumerated their potential costs and benefits. A diagram of each of the approaches to modeling the IAT we have described is shown in Fig. 3. The classical difference-score approach (left) quantifies attitudes and associations in terms of (standardized) mean differences in response times between two conditions, under the assumption that the sign and direction of this difference is proportional to a participant's degree of bias. The multinomial processing tree approach, including the Quad and ReAL models (middle left), uses accuracy on IATs to estimate the probabilities of different events happening – such as a participant's bias being being activated or overridden or a stimulus being correctly discriminated. Third, we have the diffusion decision model (center right), which quantifies both accuracy and response times within a condition in terms of the quality of incoming information (drift), response caution (threshold), bias, and non-decision processes like stimulus encoding and

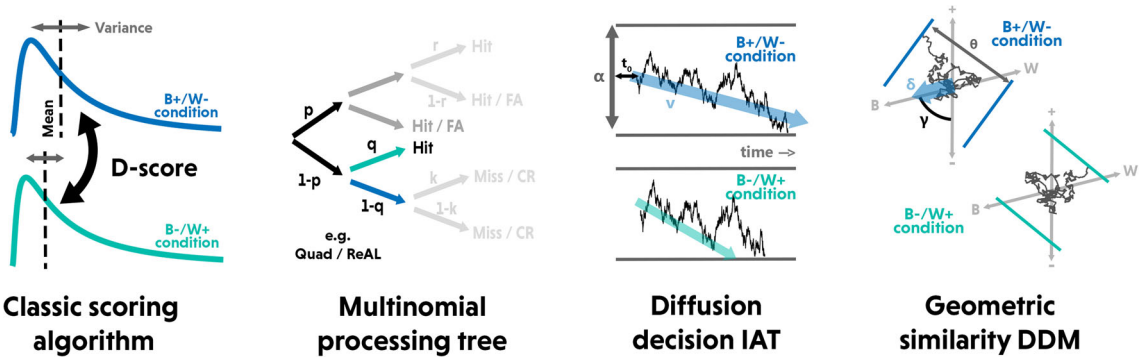

Fig. 3 A diagram of each approach to quantifying behavior on the IAT. In this paper, we compare the classic scoring algorithms like the D-score (left) as well as the diffusion decision model (as it has been previously applied; center-right) and a new model based on geometric relationships among concepts involved in performance on the IAT. While we focus on comparisons between congruent (green) and incongruent (blue) conditions here, the GSR-DDM is fit to all four IAT conditions

motor actions. Finally, we have our new geometric similarity approach (right panel), which proposes that performance across all conditions is driven by the stimulus and the degree to which it activates different category responses.

Below, we examine a model from each of these four examples. We test the *D*-score, Quad model, simple diffusion model, and geometric similarity / association model, examining their ability to predict important outcomes related to outgroup contact and motivation from individual-level parameter estimates.

There are clear issues we identified with the first three approaches that are addressed in the new model. First, classic scoring algorithms like the *D*-score, as well as the diffusion decision model, suffer from the compound variance problem, where difference scores result in much greater measurement error than single parameters alone (Bereiter, 1963; Thomas and Zumbo, 2012; Overall and Woodward, 1975; Gardner and Neufeld, 1987). Second, the three approaches on the left all ignore valid information contained in IAT data – classic scoring algorithms ignore both accuracy and non-paired (single-category) conditions, multinomial processing trees ignore response time information, and the diffusion decision model ignores (or at least does not capitalize upon) non-paired conditions.

Finally, each of the models presents an impoverished view of what participants are doing on an IAT. Any model has to deal with some degree of simplification or abstraction in order to be effective (Sun, 2008). However, ignoring processes like the speed at which different types of stimuli are processed or the degree to which participants try to disguise or control their performance across conditions omits critical information that is directly or indirectly relevant to understanding behavior. In the discussion, we revisit how the validity of a model interacts with its reliability and predictive power, emphasizing the importance of explicitly modeling relevant psychological processes as one seeks to understand latent traits and processes from behavior.

Below, we test our new approach and its ability to reliably capture behavior on several IATs. An ideal model of behavior should have high test-retest reliability to its parameters, which should be associated with other measures and real-world consequences of what they seek to quantify. At minimum, we can show that this new model out-performs current and past approaches in both respects.

Methods

To assess the performance of the model, we tested both the reliability and the validity of its parameters using two large IAT data sets for each analysis. The full GSR-DDM model capitalizes on information both about the speed of participants' responses as well as information about their accuracy, rather than ignoring one as in other modeling approaches (Meissner and Rothermund, 2013; Calanchini and Sherman, 2013). While future iterations may be fit to data that exclude accuracy or response time to deal with practical constraints imposed by existing datasets, the present version requires both accuracy and response time data to best estimate its parameters. Given the informativeness of both sources of data, we strongly encourage any researchers considering using IATs to record both accuracy and response times – an individual's data might not show an effect in accuracy or speed alone due to their ability to deliberately control the speed-accuracy tradeoff (Reed, 1973; Wickelgren, 1977; Heitz, 2014). This is made clear from diffusion modeling of IATs (Klauer et al., 2007; Röhner and Lai, 2021), where thresholds can be seen to vary across conditions of the task.

Fortunately, there are several large existing data sets that include information on both accuracy and response time that

we can use to test the reliability and validity of the estimates we obtain from the model. We focus on four main studies: two to examine the test-retest reliability of the parameters, and two to examine its ability to predict important real-world behaviors relevant to the attitudes or associations IATs seeks to measure.

These studies each included 1-2 blocks of 30-40 trials in each IAT condition, with an average of 60 trials per condition – typically 40 trials per condition for non-target condition, and 80 trials per condition for congruent / incongruent condition. As is fairly common in social psychological tasks, the number of trials is lower than most dynamic decision-making tasks for which the diffusion model is estimated, making the hierarchical constraints we use particularly important (Pleskac et al., 2018).

Transparency and openness

The goal of this study was to evaluate performance on existing data in order to evaluate whether our modeling approach improves on current methods for analyzing IAT data. We therefore used secondary data for all of the analyses presented here. Each data set was selected *a priori* by one of the authors (L.H.I.) based on a set of constraints on sample size and trial-level information provided by another author (P.D.K.). This was done to avoid a biased data-selection process whereby data were selected that might favor the new modeling approach. Ultimately, we selected two data sets with a test-retest design for reliability analyses [from] Gawronski et al. (2017) and two data sets that included relevant outcomes to assess predictive validity (Buttrick et al., 2020). To avoid any file drawer effects (Rosenthal, 1979), we report the results of all four studies regardless of the results.

The data that were used here can already be found at `osf.io/792qj` (reliability studies) and `osf.io/6d7xp` (predictive validity studies). The JAGS model code for GSR-DDM and the MATLAB code for accessing and running it are provided at `osf.io/znsfb`. We also provide JAGS code for the diffusion model and Quad model. Note that running these scripts requires JAGS, MATLAB, and the JAGS-MATLAB interface matjags (Steyvers, 2011).

Since the inferential purposes of our study were primarily *abductive* (i.e., what explanation / model best accounts for the data?) as opposed to confirmatory hypothesis testing, preregistrationg was largely irrelevant. Further, the methods we use are not sensitive to issues of multiple testing and the chosen data sets were selected for analysis a priori (Szollosi et al., 2020; Devezer et al., 2020; Rubin and Donkin, 2022; Rubin, 2020).

Assessing test-retest reliability

As we outlined above, part of the issue with accurately measuring the reliability of behavioral measures is the lack of good generative models explaining how observed data are related to latent processes. Modeling processes underlying a single test session – as in the ReAL, Quad, and diffusion models – is a step in the right direction. However, there are (at least) two levels of error that enter into assessments of test-retest reliability. Modeling helps account for *trial-level* error, or variability in response times and accuracy that occurs within a single session. However, it does not account for *session-level* error, or variability in performance that naturally occurs from day to day or session to session. Properly incorporating this error into our estimates of reliability is critical to understanding how reliable our measures actually are (Haines et al., 2021; Rouder and Haaf, 2019).

The approach that we took to re-assess reliability on IATs is shown in Fig. 4. Rather than estimating a model separately for time points 1 and 2, we used a joint model that

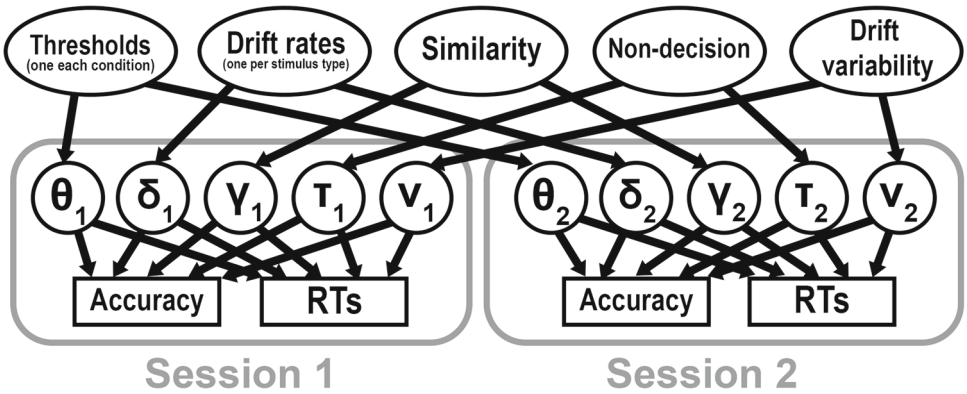

Fig. 4 Diagram of the structure of the model we used for reliability. Observed response times [RTs] and accuracy at multiple measurement time points (gray boxes) are viewed as the product of latent thresholds, drift rates, and non-decision times. In turn, we estimate the covariance structure of these latent cognitive processes across time points, accounting for error in both the observed data (RT, accuracy) and error in our measurement of the latent cognitive processes at each time point

simultaneously fit behavior across both time points using a factor-based link (Kvam et al., n.d.; Turner et al., 2017). This constitutes a shift from treating behavior at each time point as a separate, independent measurement to treating behavior as two measurements of a common set of latent cognitive processes. Formally, we fit a bivariate normal distribution that specified the relationship between parameter estimates from the first session (x) and the parameter estimates for the second session (y) in terms of their means (M_x and M_y), variances (V_x and V_y), and covariances (Σ_{xy}) alongside the session-level parameters that we normally fit with these models (conceptual-similarities γ, drift rates δ, thresholds θ, non-decision times τ, and drift variabilities ν; see Fig. 4). This accounts for both trial-level and session-level error, and allows us to control for these dual sources of error when estimating the test-retest reliability of our model parameters.

We tested this new approach to estimating reliability using a data set from a series of studies where participants completed various implicit measures at two different timepoints (Gawronski et al., 2017). Specifically, we re-assessed the test-retest reliability of Race and introversion-extraversion IATs administered twice over a one to two month interval. We compare this against both the *D*-score (Greenwald et al., 1998, 2003) and a diffusion model that is fit separately to congruent and incongruent conditions. In the original paper, the reliability of both IATs was assessed as being relatively poor, with correlations of $r = .44$ and .63 between IAT *D*-scores across the two timepoints for Race and introversion-extraversion IATs, respectively. Accordingly, the findings are commonly cited as a key piece of evidence in failing to establish the validity of IATs as valid measures of relatively stable individual differences (Gawronski, 2019; Payne et al., 2017). In the results, we show that a re-analysis modeling the measurement relationships among latent variables and time points ultimately paints a much more favorable picture of the reliability of performance on the IAT, as well as quantifying how these parameters change (e.g., with practice) across sessions.

Assessing predictive validity

In addition to reliability, we also sought to examine whether our modeling approach could provide more a valid account of behavior on IATs. By virtue of disentangling performance into multiple cognitive processes, it already improves upon the discriminant validity of previous approaches by quantifying distinct elements of performance. Likewise, modeling the underlying cognitive processes gets us closer to a complete description of performance and helps IATs align more closely with the established literature on lower-level perception and decision-making (Weber and Johnson, 2009). However, much of the debate surrounding IATs has concerned their ability to predict real-world outcomes and by extension their *predictive* validity. To test whether the model parameters were indeed better predictors of important outcomes, we fit GSR-DDM to another secondary data set (Buttrick et al., 2020) and used its estimates to predict relevant outcome measures that were related to the attitudes or associations those IATs sought to quantify. Buttrick et al. (2020) sought to compare a typical regression approach versus a structural equation modeling approaches for estimating the unique predictive validity of the IAT above and beyond analogous self-report measures. To do this, they randomly assigned participants (volunteers from the Project Implicit website; total $N > 14{,}000$) to one of ten experimental conditions where they completed IATs and self-report measures whose content was manipulated to target different social groups. Although their study included 10 pairs of social groups as well as self-reported criterion measures spanning across five criterion domains, we selected two IATs that are both widely used in the extant research literature and socially relevant (Race IAT and Sexuality IAT) along with self-report measures for two highly relevant criteria (internal motivation to respond without prejudice and prior intergroup contact).

The Internal Motivation to Respond without Prejudice scale was originally developed to understand how internal and external motivation influenced people's race-related attitudes and behaviors (Plant and Devine, 1998). It is widely used in the IAT literature and has been adapted to other contexts including prejudice toward gay men and lesbians. The five-item internal motivation subscale is associated with relevant outcomes in both domains, and it is especially relevant to implicit measures like the IAT because it is theorized to index a relatively automatic process. For example, internal motivation is associated with lower levels of implicit and explicit race bias (Devine et al., 2002), increased automatic activation of egalitarian goals (Johns et al., 2008), and positive interracial interactions LaCosse and Plant (2020). Similarly, it is associated with positive sexuality attitudes (Ratcliff et al., 2006), more positive experiences when interacting with gay men (Lemm, 2006), and greater effectiveness of diversity training (Lindsey et al., 2015).

The race and sexuality prior interpersonal contact measures each include five items adapted from commonly used items in previous research (Pettigrew and Tropp, 2006) and are especially relevant because it is the available criterion measure that most closely approximates actual real-world behavior. Although they were modified and aggregated ad hoc to be administered on the Project Implicit demonstration website, the full sets of race and sexuality items have respectively been found to be associated with negative racial outgroup and positive ingroup evaluations (Rae et al., 2020) and with less individual- and contexual-level prejudice toward gay men and lesbians (MacInnis et al., 2017).

Results

All of the model fits presented in the results were generated using hierarchical Bayesian methods, which estimate not only the individual-level parameters that characterize individual differences in cognitive processes but also the group-level central tendency of these parameters (Shiffrin et al., 2008). In the next section, we examine a neural network-based approach to fitting the model. Such an approach is more accessible in that it can be embedded into point-and-click model fitting tools, but it does not estimate the covariance across testing sessions (and thus is not as useful for estimating reliability) like the Bayesian joint model shown in Fig. 4. Therefore, we focus for now on the Bayesian methods. Bayesian analyses compute an approximate posterior distribution of parameter values, which assigns probabilities to different possible values of the parameters of the model. For simplicity and brevity, we typically report the mean value of each parameter along with the 95% highest density interval [HDI], which specifies the range of the 95% most likely values of each parameter. It is analogous to a confidence interval except it directly quantifies the most likely values for a parameter as opposed to quantifying the range in which we would expect them to fall if the sampling process were repeated many times, thus providing an overall more coherent and interpretable measure of uncertainty (Kruschke, 2014; Kruschke and Liddell, 2018).

Unless otherwise specified, these values were computed using a Gibbs sampler [JAGS] Plummer (2003) using 4 chains of 5000 samples each, with 1000 burn-in samples per chain. In all cases, these chains converged according to both visual inspection and r-hat statistics for convergence [all $\hat{r} < 1.001$;] Gelman and Rubin (1992); Roy (2020).

Reliability

The results of the first study (Self-Concept / introversion-extraversion IAT), Study 1a from Gawronski et al. (2017) are shown in Fig. 5. The top panels show the estimates of the model parameters from the first testing session (x), compared against the estimates from the second testing session (y). There are a few key findings to note here. We discuss the mean estimates of model parameters in greater detail in the Predictive Validity section, but there are several findings specifically related to reliability. First, GSR-DDM model parameters related to the decision process – specifically, similarity, thresholds, and drift rates – showed a high degree of test-retest reliability, with all linear correlations at least .77. Non-decision time has somewhat lower reliability at only .51, but this parameter specifically indexes non-decision components of response time and is therefore not of particular theoretical interest; its lower reliability is not particularly surprising, as it is a catch-all parameter that quantifies multiple different processes like stimulus encoding and response execution.

In addition to the reliability of the parameters, we also observed that the base drift rate – indicating how well participants are able to assign words or faces to categories regardless of the associations between categories on either side of the screen – were higher in the second testing session than in the first. This is exactly what we would expect from practice effects, as participants get faster and more accurate as they do the task more. Critically, drift rates were the only parameter that changed between sessions – not similarity. This means that the model is capturing one process related to decision-making that is stable across time (associations) and another that improves with practice (drifts). This lends credibility to the conceptual-similarity parameter as a measure of latent associations that participants have among concepts, and suggests that the model is showing a high degree of discriminant validity by disentangling practice effects from core individual differences. This stands in contrast to unitary measures of performance like the *D* score or even advanced approaches like the diffusion model where associations and information processing speed are both combined into a single measure of drift.

In the model, rather than merely correlating the estimates from the first and second testing session (as in the top panels of Fig. 5), we estimated the reliability of each parameter directly by estimating the variance-covariance matrix for parameter values across the two sessions. The covariance is estimated in a Bayesian way that obtains a posterior distribution describing the likelihoods of different values for the reliability given the data (Haines et al., 2021). Results from this analysis are shown in the bottom panels of Fig. 5. The approach was stricter than the simple correlations due to the inclusion of a prior centered at zero, and thus resulted in lower mean estimates of reliability relative to the top panels. Bayesian analyses require priors, and in all cases we strove for relatively vague ones (i.e., ones that would not favor particular conclusions a priori). However, it is clear that almost all of the model parameters – association strength (estimated reliability $M(r_\gamma) = .77$, posterior 95% HDI = [.72, .85]), threshold ($M(r_\theta) = .71$, 95% HDI = [.67, .75]), and drift for identity-related words ($M(r_{\delta_I}) = .90$, 95% HDI = [.73, .99]) and intro/extraversion related words ($M(r_{\delta_E}) = 0.67$, 95% HDI = [.56, .79]) – show greater test-retest reliability than the IAT *D* score, shown as a vertical dotted black line.

Compared to the reliability of the traditional diffusion model contrasts, these reliabilities are exceptionally high, as shown by the dashed lines in Fig. 5. Ironically, the raw drift rates of the diffusion model show quite high reliability ($r(\delta_I) = .67$ and $r(\delta_E) = .73$), as do the thresholds ($r_(\theta_C) = .64$ and $r_(\theta_I) = .66$). By taking the difference, their error variability is doubled and the metric becomes unre-

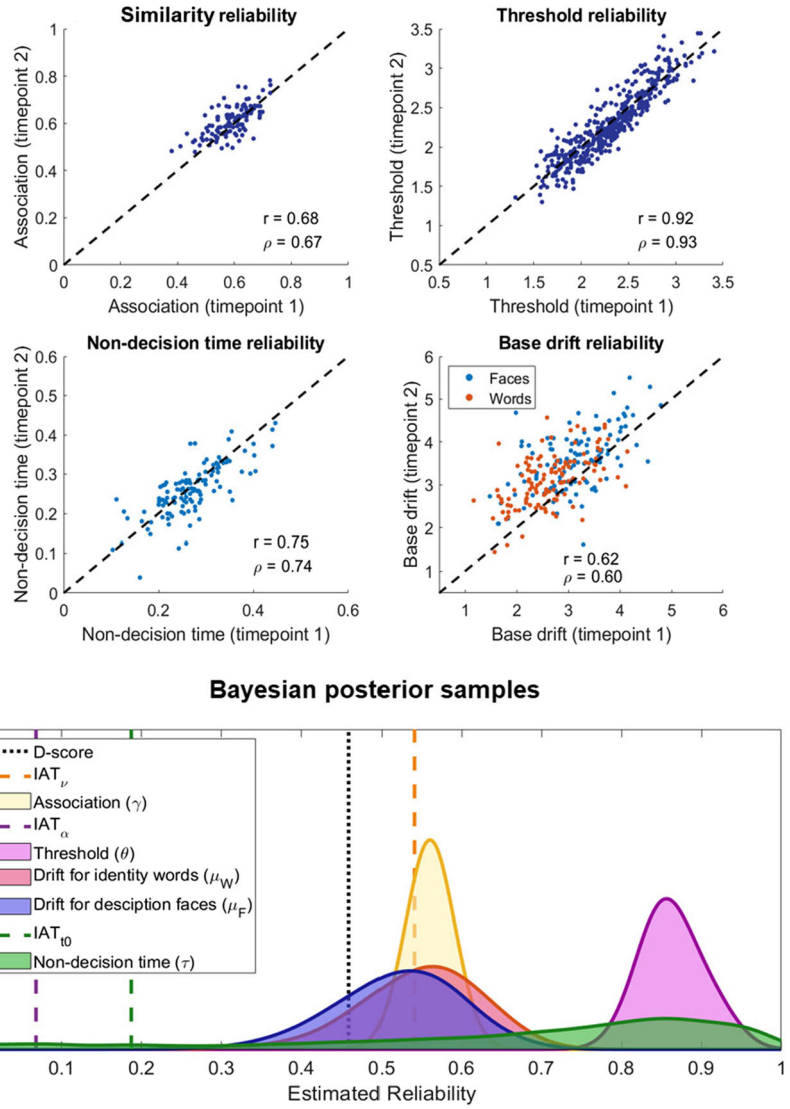

Fig. 5 Estimates of model parameters (top) and the posterior distribution of their reliability (bottom) on the introversion-extraversion IAT. In the top panels, the estimated values of each parameter for the first test session (x) and second test session (y) are shown. In the bottom panels, posterior distributions of estimated test-retest reliability for each of the model parameters are shown. These are compared against the test-retest correlations of the IAT D-score (black dotted line) and the differences between congruent and incongruent conditions for drift rates (IAT_v, orange dashed line), thresholds (IAT_a, purple dashed line), and non-decision time (IAT_{t0}, green dashed line) from a traditional diffusion model

liable ($r(IAT_v)$ =.61, $r(IAT_a)$ =.11, and $r(IAT_{t0})$ =.03; colored dashed vertical lines in Fig. 5), clearly illustrating the issue of compound error in difference scores. This is exactly the reason we approached the problem using one similarity parameter to create differences between congruent and incongruent conditions as opposed to computing a contrast coefficient.

The results of the second reliability analysis using the Race IAT [Study2b] Gawronski et al. (2017) are shown in Fig. 6. Like the introversion-extraversion IAT, it showed relatively high reliability of the model parameters relative to the *D*-score. However, the reliability of the similarity parameter ($M(r_\gamma) = .56$, 95% HDI = [.55, .58]) and the drift parameters (Faces: $M(r_{\delta_F}) = .55$, 95% HDI = [.40, .69]; Words: $M(r_{\delta_W}) = .52$, 95% HDI = [.37, .67] was somewhat lower than in the introversion-extraversion IAT analysis. The difference in association strength reliability appears to stem primarily from smaller overall variance of the estimates. That is, people appear to have less variability in their associations between race and positive/negative words than they do between their representations of self/others and particular traits.

Conversely, the test-retest reliability of the threshold parameters in this study was much higher ($M(\theta) = .86$, 95% HDI = [.82, .93]). Even if participants behavior in terms of information processing was more variable, the level of caution they implemented across testing sessions was highly consistent, to the point of being comparable to the reliability of some trait-level measures. As in the introversion-extraversion IAT, these reliabilities substantially out-performed the *D*-score ($r_D = .46$) and were on par or greater than the reliability of mean response times ($r_{RT} = .55$).

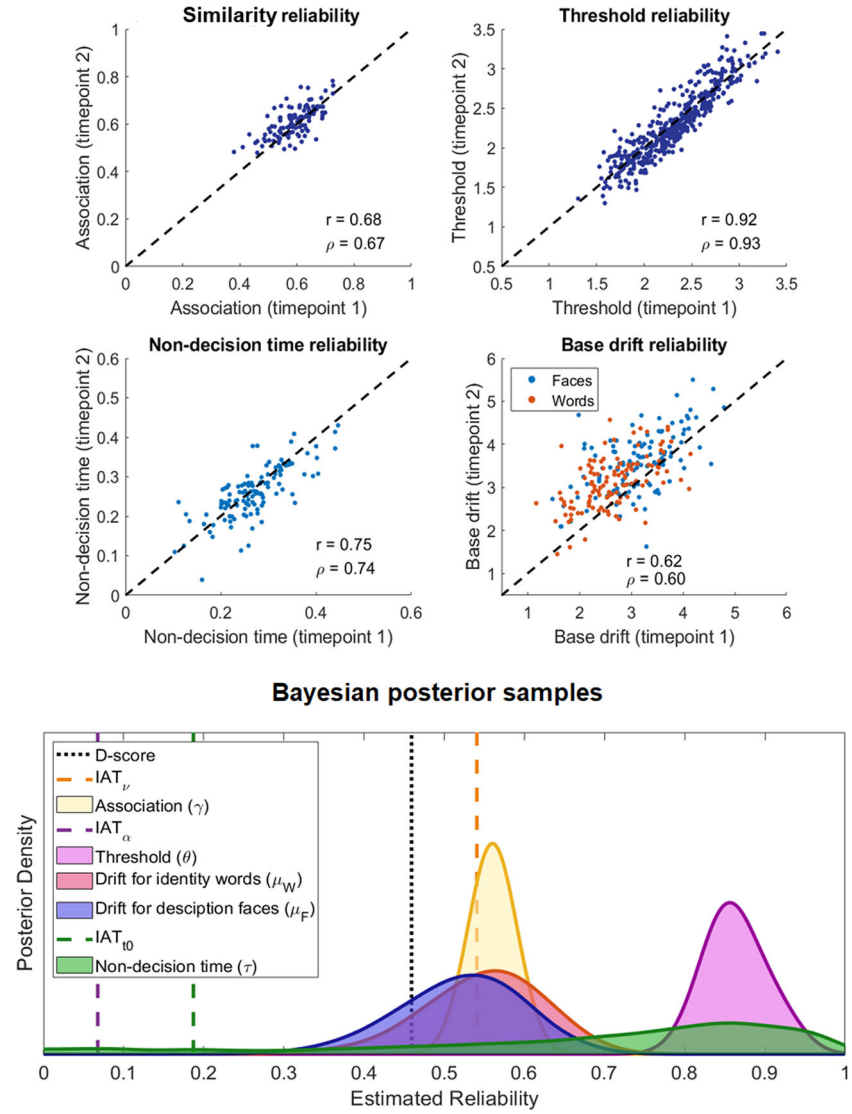

Fig. 6 Estimates of model parameters (top) and the posterior distribution of their reliability (bottom) on the Race IAT. In the top panels, the estimate values of each parameter for the first test session (x) and second test session (y) are shown. In the bottom panels, posterior distributions (colors) of estimated test-retest reliability for each of the model parameters are shown. These are compared against the test-retest correlations of the IAT D-score (black dotted line) and the test-retest correlation of differences between congruent and incongruent conditions for drift rates (IAT_v, orange dashed line), thresholds (IAT_a, purple dashed line), and non-decision time (IAT_{t0}, green dashed line) from a traditional diffusion model

The test-retest reliability of nondecision time was highly uncertain, $M(r_\tau) = .46$ (95% HDI = [−.49, .99]), preventing any clear conclusions regarding the reliability of this parameter. As before though, the most important parameters out-performed the diffusion model contrasts ($r(IAT_v)$ =.53, $r(IAT_a)$ =.07, and $r(IAT_{t0})$ =.19; colored dashed lines in Fig. 6). Also as before, the diffusion model could reach higher reliability by foregoing contrasts, as the individual condition drift rates had fairly high reliability ($r(\delta_F)$ =.70 and $r(\delta_W)$ =.69) as did the thresholds ($r(\theta_C)$ =.54 and $r(\theta_I)$ =.51).

A particularly interesting finding related to the drift rates is that the face stimuli (M(μ_F) = 3.34, 95% HDI = [3.07, 3.62]) appeared to be processed faster than the words ($M(\mu_W) = 2.87$, 95% HDI = [2.59, 3.15]). Holistic visual processing of images is often found to be an efficient process relative to serial or lexical processing (Richler and Gauthier, 2014), so this appears to reflect real differences that we might have expected a priori. We elaborate on this finding further in the latter part of the next section.

Predictive validity

It is clear from the reliability analyses that the test-retest reliability of the GSR-DDM model parameters exceeds that of simple metrics of performance like the *D*-score. In theory, this should translate to greater predictive validity. Observed correlations between constructs r_{xy} are determined by both the "true" relationship between the constructs (ρ_{xy}) as well as the reliability of the predictor(s) r_{xx} and the reliability of the outcome r_{yy}. As a result, if we can improve the reliability of our predictor (r_{xx}), we should be able to better predict any outcome (y) provided the true relationship between predictor and outcome does not change by using a slightly different version of our predictor (x). Put simply, being better able to

measure individual differences in cognitive processes should mean that we can better predict other outcomes.

For both studies, GSR-DDM was fit in a hierarchical Bayesian way as in the previous studies. It included nine total free parameters: baseline drift for face stimuli, baseline drift for word stimuli, an association parameter indexing the relative degree of association between White/Straight or Black/Gay and positive or negative words, four thresholds for the four conditions, a drift variability parameter to account for slow errors, and the non-decision time. These nine parameters were used as predictors of internal motivation and contact outcomes, and compared against the *D*-score (a single predictor) as well as the difference scores from the diffusion model (IAT_v, IAT_a, and IAT_{t0}) and the parameters of the Quad model. The GSR-DDM having more parameters makes it more complex, but ultimately serves as a benefit to the model as a whole by indexing multiple cognitive processes that can predict real-world outcomes. For example, a desire to manage one's impressions could appear as differences in either θ_C or θ_I, which may predict responses on an explicit self-report, while differences in similarity γ may predict behavioral or self-report outcomes that cannot be as easily controlled through impression management (Röhner and Ewers, 2016).

Model fit

There are several ways to assess predictive validity of the GSR-DDM model parameters relative to the *D*-score. If we simply look at total variance in the outcomes that each one can account for, the GSR-DDM is the clear winner (middle column of Table 1. However, in some cases the GSR-DDM may perform better simply because it is more complex (more parameters and thus more predictors). To control for greater number of predictors in GSR-DDM, we penalized model fit for each parameter used to predict outcomes. We tested two different metrics that favor models with fewer predictors: a classical measure called Adjusted R^2 (Shieh, 2008; Yin and Fan, 2001) and a Bayesian measure called the deviance information criterion [DIC] (Spiegelhalter et al., 2014).

Even after applying the correction to the model fit indices, GSR-DDM still out-performed the *D*-score, Quad model, and diffusion contrasts on nearly every outcome measure, as shown in the DIC and Adjusted R^2 columns of Table 1. The only model to ever edge out the GSR-DDM, the Quad model, did so only on one metric (Adjusted R^2), and was inferior to the GSR-DDM when predicting every other outcome. Note that lower DIC scores indicate better predictive validity, with differences of at least 10 between models indicating strong support for the better-performing model (Spiegelhalter et al., 2002; Schwarz, 1978). GSR-DDM exceeds this criterion relative to the *D* score in all but one of the outcome measures (Contact on the Race IAT), where it only improves the DIC by 2. Overall, GSR-DDM shows clearly superior performance relative to the *D*-score, as well as the simple diffusion model and Quad model, in predicting contact and motivation outcomes.

It is also worth examining which model parameters pull the most weight when predicting the outcomes of interest. The estimates of the relationship between the motivation

Table 1 Model fit metrics for the *D*-Score, GSR-DDM, diffusion model, and Quad model for each of the studies and outcomes we examined

Study	Outcome	Model	DIC	R^2	Adjusted R^2
	Motivation	D-score	4520	0.00223	0.00161
		GSR-DDM	**4503**	**0.0206**	0.0151
		Diffusion	4520	0.00234	0.0005
		Quad	4504	0.0188	**0.0163**
Race					
	Contact	D-score	4477	0.029	0.0283
		GSR-DDM	**4475**	**0.0406**	**0.0352**
		Diffusion	4476	0.0298	0.0279
		Quad	4494	0.0164	0.0139
	Motivation	D-score	4194	0.0245	0.0239
		GSR-DDM	**4161**	**0.0801**	**0.0745**
		Diffusion	4182	0.0304	0.0284
		Quad	4204	0.0171	0.0145
Sexuality					
	Contact	D-score	4150	0.0526	0.052
		GSR-DDM	**4099**	**0.0957**	**0.0902**
		Diffusion	4140	0.058	0.0561
		Quad	4187	0.0281	0.0255

and contact outcomes (for the Race and Sexuality IAT) and each model parameter, as well as the *D*-score, Quad model parameters, and the three diffusion model contrast scores, are shown in Table 2. The conceptual-similarity parameter was most strongly related to each outcome, which is intuitive given it was designed specifically to quantify the similarity among cognitive representations of race, sexuality, and valence (positive / negative). In fact, this parameter alone out-performed the *D*-score, Quad parameters, and diffusion contrasts on almost every outcome measure. There are only two exceptions, which are the D (discriminability) parameter in the Quad model for Race-Motivation outcome and the *D*-score for the Sexuality-Contact outcome. Critically, although the improvement in prediction moving from the *D*-score to the similarity parameter is consistent, the small overall improvement in predictive power with this parameter alone is complemented by significant predictive power arising from the other parameters in the GSR-DDM such as the thresholds. Together, the parameters of the GSR-DDM constitute a considerable step up from the other existing models, as shown in Table 1.

Those interested in a more "process-pure" measure of similarity than the *D*-score need look no further than the similarity parameter, which contains less noise / error due to GSR-DDM's ability to disentangle associations from effects related to thresholds, drift rates, and other parameters. These findings suggest that the strength of the model is not only its greater number of estimable predictors, although this is certainly a strength in terms of discriminant validity, but it can also isolate the impact of associations from those of other cognitive processes involved in performance on IATs.

In addition to the conceptual-similarity parameter, the threshold for the incompatible block trials (e.g., with Gay people+positive vs. Straight people+negative) adds considerable predictive power. It seems that participants who set higher thresholds in this condition had lower motivation and contact scores. This could be because participants are trying to manage their impressions in these conditions and avoid bias-indicative mistakes by lengthening their response times, as with participants who try to fake their IAT scores (Röhner et al., 2013). It could also be that participants who are more biased are simply attuned to the fact that this condition could be more difficult or more sensitive, and adjust their thresholds based on this perceived difficulty. In either case, it is clear that both similarity / interference and response caution in the face of bias-induced conflict (measured by threshold adjustment) are predictive of participants' interactions with minoritized group members (the majority of whom form an outgroup to participants).

Beyond these parameters, there are remaining significant correlations between base drift rates and motivation / contact outcomes. These positive correlations indicate that participants who generally performed better on an IAT (were more accurate, made faster responses) also reported more internal

Table 2 Ability of the *D*-score, diffusion model, Quad model, and GSR-DDM model parameters (rows) to predict motivation and contact outcomes. Values in bold indicate significant / credible predictors whose HDIs exclude zero

	Race				Sexuality			
	Motivation		Contact		Motivation		Contact	
Predictor	Mean	95% HDI	Mean	95% HDI	Mean	95% HDI	Mean	95% HDI
D-score	-0.05	[-.10, .00]	**-0.17**	**[-.22, -.12]**	**-0.16**	**[-.21, -.11]**	**-0.23**	**[-.28, -.18]**
IAT_v	-0.04	[-.14, -.04]	**-0.14**	**[-.19, -.09]**	**-0.08**	[-.14, -.04]	**-0.13**	**[-.18, -.08]**
IAT_a	0.02	[-.03, .07]	**0.08**	**[.04, .13]**	**0.13**	**[.08, .18]**	**0.17**	**[.12, .22]**
IAT_{t0}	0.03	[-.02, .08]	0.03	[-.01, ,.08]	0.02	[-.03, .07]	**0.07**	**[.03, .12]**
D	**0.15**	**[.08, .22]**	.08	[-.00, .16]	**.09**	**[.01, .16]**	**.11**	**[.03, .18]**
AC	.02	[-.03, .07]	**.10**	**[.05, .15]**	**.06**	**[.00, .11]**	**.10**	**[.05, .15]**
OB	-.04	[-.11, 0.03]	-.06	[-.14, .02]	.01	[-.07, .09]	-.03	[-.10, .05]
G	.05	[-.00, .10]	.01	[-.03, .07]	.03	[-.02, .08]	.07	[-.02, .12]
Base drift (faces)	0.02	[-.02, .06]	0.00	[-.03, .04]	**0.09**	**[.04, .13]**	**0.07**	**[.02, .11]**
Base drift (words)	**0.12**	**[.05, .18]**	0.05	[-.02, .12]	-0.04	[-.11, .03]	0.04	[-.03, .10]
Similarity	**-0.07**	**[-.12, -.01]**	**-0.18**	**[-.22, -.12]**	**-0.17**	**[-.22, -.12]**	**-0.21**	**[-.26, -.16]**
Threshold (faces only)	0.04	[-.09, .02]	**0.07**	**[.02, .13]**	-0.03	[-.09, .02]	-0.03	[-.09, .02]
Threshold (words only)	0.00	[-.06, .06]	-0.01	[-.07, .05]	-0.01	[-.07, .05]	0.07	[.02, .13]
Threshold (compatible)	0.02	[-.04, .08]	0.04	[-.02, .10]	0.04	[-.01, .10]	0.06	[.00, .13]
Threshold (incompatible)	0.01	[-.05, .06]	**-0.08**	**[-.13, -.02]**	**-0.14**	**[-.19, -.08]**	**-0.20**	**[-.25, -.14]**
Drift variability	**0.07**	**[.01, .14]**	-0.03	[-.09, .04]	0.04	[-.04, .11]	0.05	[-.02, .11]
Non-decision time	-0.01	[-.98, .91]	0.00	[-.97, .92]	0.01	[-.95, .94]	0.00	[-.97, .92]

motivation to be unbiased as well as more prior contact with outgroup members. Drift rates have been increasingly viewed as global measures of neural processing speed (Schubert et al., 2019) and intelligence (Lerche et al., 2020; van Ravenzwaaij et al., 2011a), indicating perhaps that more intelligent or better educated participants tend to be less biased in their behaviors toward others. This interpretation is consistent with findings that executive functioning is associated with smaller IAT effects and less explicit bias (Klauer et al., 2010; Ito et al., 2015).

Put together, the conceptual-similarity parameter alone provides an incremental advantage over existing approaches. Even if we were to ignore all other parameters entirely, this would be a small victory for the GSR-DDM. However, when we consider the other parameters of the model and the predictive power that they confer over and above similarity, alongside the conceptual benefits of disentangling these cognitive processes, the result overwhelmingly supports the GSR-DDM on every outcome and data set we have tested.

Characterizing the model outcomes

Given that this is the first time this type of model has been applied to such large IAT data sets, it is prudent to investigate the patterns of GSR-DDM parameter estimates for each data set. This can help us understand more about issues related to biases that exist in large populations, how participants perform with different stimuli, and how they generally shift their thresholds across conditions. It is rare that dynamic cognitive models are fit to such a huge data set, as typical experiments encompass only a few participants for thousands of trials each rather than thousands of participants for a few trials each. The use of hierarchical Bayesian approaches for model fitting were therefore particularly important here (Shiffrin et al., 2008), as each participant only had a few trials from which to estimate their parameters.

Once the model is fit, we can explore the group-level distributions of estimates for each parameter to ensure (a) that these distributions are sensible with respect to what they are theorized to measure, and (b) to explore any patterns in the data that allow us insights about the population. In general, the approach we used for model fitting appears to have turned up sensible results, indicating that the GSR-DDM is capturing realistic patterns of individual differences. Distributions of model parameter estimates across individuals for the Race IAT in Buttrick et al. (2020) are shown in Fig. 7. For each of the parameters we discuss, we report the mean group-level estimate (best estimate of the population mean) and the 95% highest density interval [HDI] on this mean. This means that the 95% HDI is not describing the distribution of

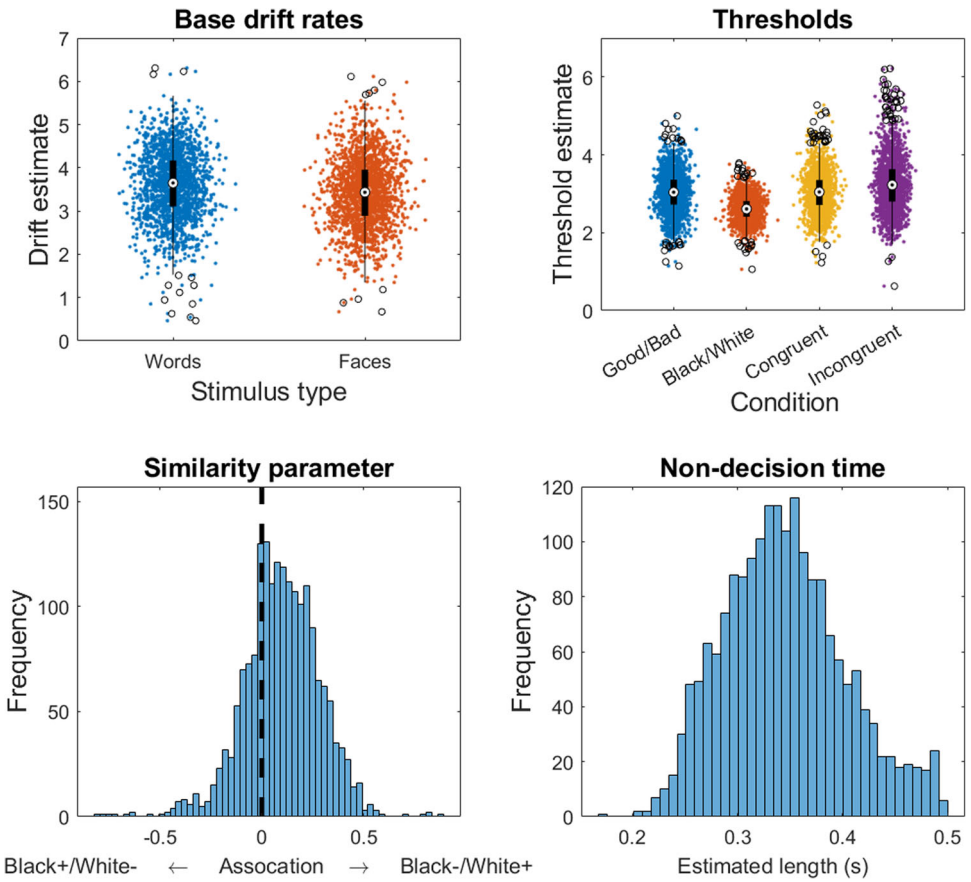

Fig. 7 Distributions of parameter estimates across all participants in the Race IAT dataset

individual-level estimates, which are shown in Figs. 7 and 8, but the uncertainty about their central tendency. Because of the immense volume of data in both experiments, these HDIs are very thin. Any comparisons between conditions with non-overlapping intervals will be significant by essentially any classical or Bayesian inferential test.

There are a few key findings here to note. First, most people are biased toward pro-White / anti-Black associations, on average. A total of 71.97% of conceptual-similarity parameter estimates lie above zero ($M(\gamma) = .10$, 95% HDI = [.09, .10]), indicating a greater similarity between White people and positive / Black people and negative than Black people and positive / White people and negative. This is consistent with findings based on the *D* score (Greenwald et al., 2003; Nosek, 2007) as well as simpler metrics like mean response times (Greenwald et al., 1998) and accuracy (Calanchini and Sherman, 2013; Calanchini et al., 2014). Out of an abundance of caution on our part, the priors for the association parameter were actually centered at zero. This means that the model likely slightly underestimated the group-level mean of the association parameter, although with over 14,000 participants, this underestimation should be negligible. Regardless, the model reproduces the classic bias findings related to race, although it does not make any claims about whether these associations or biases are "implicit" or "automatic" in any of the ways in which that term has been used (Gawronski et al., 2022; Moors and De Houwer, 2006).

As in the reliability experiment, the drift rates for faces ($M(\delta_{Faces} = 3.62$, 95% HDI = [3.53, 3.68]) were higher than the drift rates for word stimuli ($M(\delta_{Words} = 3.43$, 95% HDI = [3.34, 3.49]). This appears to reflect greater processing speed for stimuli that are processed holistically (Richler and Gauthier, 2014), and emphasizes the importance of differentiating between types of stimuli that can be presented from trial to trial in IATs.

The model does suggest that people have at least some awareness or recognition that it is more difficult to respond quickly and accurately to the incompatible block trials, as shown in the threshold estimates in the upper-right panel of Fig. 7. Participants set higher thresholds in the incompatible block ($M(\theta_{Incompatible}) = 3.25$, 95% HDI = [3.17, 3.30]) than in the race-only block ($M(\theta_{Race}) = 3.04$, 95% HDI = [2.97, 3.09]) and compatible block ($M(\theta_{Compatible}) = 3.05$, 95% HDI = [2.97, 3.10]), which are higher in turn than the valence-only block ($M(\theta_{Valence-only}) = 2.61$, 95% HDI = [2.56, 2.65]). It appears that people are more careful when they know race is involved in their choices, either because they are aware of their biases or because they know that their decisions during the IAT are meant to reflect self-relevant evaluative information with respect to a sensitive topic.

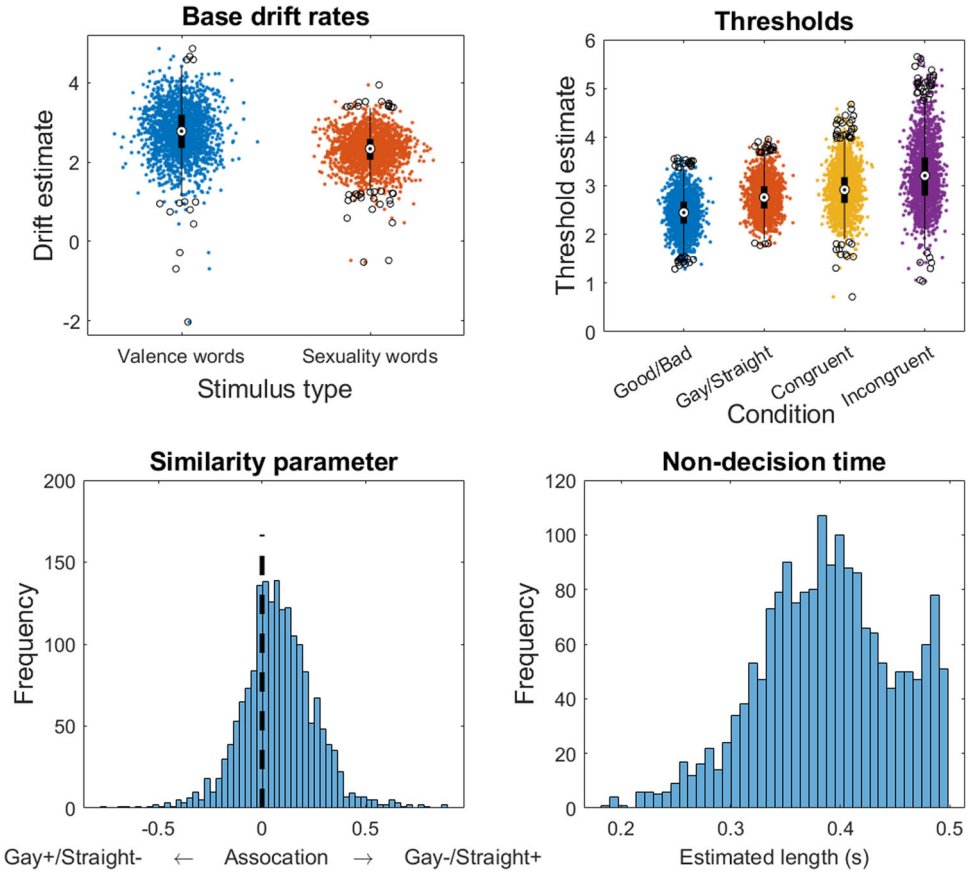

Fig. 8 Distributions of parameter estimates across all participants in the Sexuality IAT dataset

Estimations of non-decision time across participants resulted in a wide distribution of individual differences, characterized by a mean estimate of $M(\tau) = 345ms$ (95% HDI = [344, 346]). The drift rate variability indicating trial-to-trial differences in stimulus processing speed showed a relatively modest degree of drift variability with a mean of $M(\nu) = 0.53$ (95% HDI = [0.51, 0.57]).

The distributions of parameter estimates for the Sexuality IAT are shown in Fig. 8, and share many patterns of results with the Race IAT distributions. Participants in this task were slightly less biased in terms of the distribution of their conceptual-similarity parameters, with 69.30% of participants showing a Gay people/negative and Straight people/positive association ($M(\gamma) = .08$, 95% HDI = [.07, .09]). They showed slightly faster processing for valence words ($M(\delta_{Valence} = 2.77$, 95% HDI = [2.74, 2.80]) than for sexuality words ($M(\delta_{Sexuality} = 2.32$, 95% HDI = [2.29, 2.34]). There are several reasons this could be the case. For example, the valence words might be shorter, more common, or otherwise faster to encode or recover their meaning as in the introversion-extraversion IAT reported in the Reliability sections (Scarborough et al., 1977; Polich and Donchin, 1988).

Similar to the Race IAT, participants showed the highest thresholds in the incompatible block ($M(\theta_{Incompatible}) = 3.22$, 95% HDI = [3.19, 3.25]), followed by the compatible block ($M(\theta_{Compatible}) = 2.92$, 95% HDI = [2.89, 2.95]) and sexuality-only block $M(\theta_{Sexuality-only}) = 2.77$, 95% HDI = [2.75, 2.79]), with the lowest thresholds in the valence-only condition ($M(\theta_{Valence-only}) = 2.44$, 95% HDI = [2.42, 2.47]). This indicates that they had some feeling or knowledge that the incompatible block would be more difficult and, consequently, exercised greater caution and exerted more control (higher thresholds) when sexuality associations were being tested in those blocks of the IAT. Put together, both IATs indicate that performance is driven by both differences in conceptual similarity and in proactive response caution. It may be that there are elements of the biases that participants have some awareness about, or aspects that are potentially beyond participants' awareness, or alternatively that participants are simply being more careful in some conditions because they are aware it assesses sensitive material. Accounting for both clearly improves the reliability of IATs, making modeling all the more important to interpreting performance on the task.

An automated modeling tool

Despite the benefits of the modeling approach we have outlined above, modeling approaches like these remain enigmatic to many researchers. We suspect this will be a major barrier to its widespread use. While most solutions to this problem center around systemic issues like quantitative and computational training, it is also possible for modelers to make their models more accessible to a wide audience. We seek to accomplish this by using a new approach to modeling using neural networks (Radev et al., 2020a, ?, 2021; Lueckmann et al., 2019; Gutmann and Corander, 2016; Fengler et al., 2020; Cranmer et al., 2020; Sokratous et al., 2022). In this approach, rather than requiring a user to use a modeler's code to re-run their model on a new data set, a modeler instead trains a neural network to map input data (e.g., accuracy and response times) onto the most likely parameter estimates. This is made possible by the capacity of neural networks to approximate functional relationships, such as the

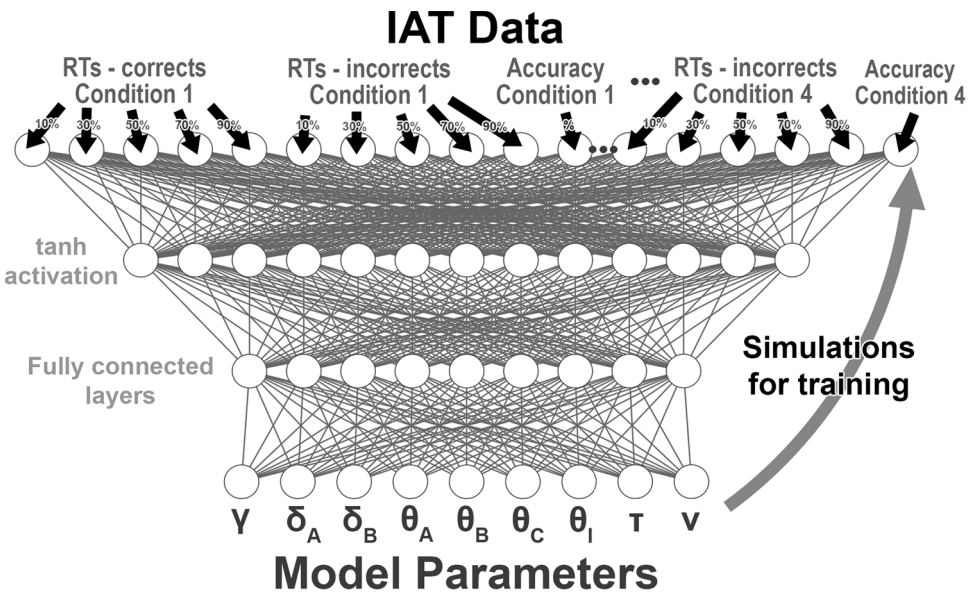

Fig. 9 Diagram of the structure of the neural network fitting approach. During training, data is simulated from the model and used to teach the network the relationship between observed IAT data (response times, accuracy) and underlying model parameters. Once trained, real IAT data can be fed into the network so as to obtain appropriate parameter estimates for that data set

relationship between model parameters and behavioral data [see the Universal Approximation Theorem] Cybenko (1989); Zhou (2020).

We implemented the GSR-DDM model in this way by training it to map behavior on an IAT onto the parameters of the model. A diagram of the approach is shown in Fig. 9. First, a modeler simulates a large volume of data from the model they want to fit – in our case, we used 100,000 simulated "participants." Each simulated participant has a true underlying set of 9 model parameters, including the similarity parameter γ; two drift rates δ_A and δ_B signifying different types of stimuli; four thresholds θ_A (first binary condition), θ_B (second binary condition), θ_C (congruent condition), and θ_I (incongruent condition); non-decition time τ; and drift variability ν. For a specific combination of these parameters, we can simulate a simulated participant's performance on the IAT, including 40-60 response times in each of the four conditions. This allows us to understand how the values of the model parameters are related to behavior – the neural network is designed to invert the simulation process by taking observed behavior and mapping it backward onto model parameters. It is enabled by using a large volume of simulated participants with known values for the different parameters, which is used to train the network.

The data set of 100,000 simulated participants was created by randomly varying the values of the 9 parameters and drawing a new data set (performance on an IAT) for each combination. The values of the parameters were each drawn from a distribution as specified below. The distribution from which each of these parameters is drawn essentially constitutes a prior distribution for the network over what it considers to be reasonable parameter values, as the relative frequency of parameter values in the training set will bias the values that the network produces. In extreme cases where there is no data, the network will simply predict the mean of this training distribution – as we would like, as the group-level mean is the best estimate one can give in absence of individual-level data. We chose the following values to simulate data from, based in part on the posterior estimates from the predictive validity study above:

$$\begin{aligned}
\gamma &\sim Beta(3, 3) \\
\delta_A &\sim Gamma(4, .8) \\
\delta_B &\sim Gamma(4, .8) \\
\theta_A &\sim Gamma(5, .75) \\
\theta_B &\sim Gamma(5, .75) \\
\theta_C &\sim Gamma(5, .75) \\
\theta_I &\sim Gamma(5, .75) \\
\tau &\sim Gamma(1.5, .25) \\
\nu &\sim Gamma(1.5, .5)
\end{aligned} \tag{1}$$

The data was fed into the network by summarizing the response time distribution in terms of 5 quantiles (10%, 30%, 50%, 70%, and 90%) for correct and incorrect responses in each condition and the accuracy in each condition, for a total of 44 inputs (5 quantiles × 4 conditions × 2 correct/incorrect + 4 accuracies). If there were no incorrect responses, zeros were passed for the incorrect quantiles. For example, if there were 40 responses in each condition of an IAT and all responses were correct, we would take the 4th, 12th, 20th, 28th, and 36th fastest responses in that condition as inputs to the network along with five 0s for the incorrects and 1.0 as the accuracy. These quantiles and accuracy statistics allowed us to summarize performance on an IAT in a way that was sufficient to identify different values of the model parameters. We also tried versions of the network where we fed in all 160 (for 40 trials / condition) or 240 (for 60 trials / condition), but this approach would not work if there was any missing data. We also tried versions where a kernel density estimator was passed over the response times to approximate a probability density function (Turner and Sederberg, 2014; Holmes, 2015) before passing the probability densities as inputs to the network, but performance was no better than the current approach.

The structure of the neural network was designed to condense the information in the IAT data down into information in the parameters. To do so, it helps to have multiple layers decreasing in size, allowing the network to iteratively condense its representation of the inputs into ones that are closer in dimensionality to the outputs (Jin et al., 2021). Specifically, we decreased the size of each successive hidden layer by 50-75% (Walczak and Cerpa, 1999; Stathakis, 2009), going from 44 to 25, 15, and then 9 nodes in each layer, with the final layer feeding into a regression layer to predict the generative model parameters. We tested deeper (more layers) and wider (more nodes per layer) up to 100 nodes × 5 layers, but there was not a substantial improvement in network performance with either of these manipulations. Fewer nodes often decreased performance, so although they improved fitting time, we persevered with the original network structure.

In addition to the 100,000 simulated data sets used to train the network, we generated an additional 100,000 simulated data sets that were used as a validation data set. The trained network was fit to these simulated data as an out-of-sample prediction, allowing us to check for overfitting and other issues that can arise with neural network-based approaches. For the training and validation sets, we estimated the parameters based on each set of inputs and compared predicted parameters from the neural network to the true parameters that were used to generate the data.

Results

A comparison between the true and estimated parameters for both the training set (blue) and the validation set (orange) is shown in Fig. 10. In general, the recovery of the true parameters was excellent for both the training set and validation set, with non-decision time and similarity slightly worse than the other parameters. There was no difference for any parameters between estimation for the training set and estimation for the validation sets, indicating that the network is free of overfitting. Therefore, we only report the correlation between true and estimated parameters in Fig. 10.

The performance of the network in recovering known parameter values provides evidence of both the model's ability to be recovered (a non-trivial component of model design) and the neural network's ability to carry out the parameter estimation. Fundamentally, it means that if the model nearly enough approximates the true structure of the cognitive processes underlying performance on an IAT, it should be able to capture the values of these processes with reasonable fidelity. Since there were no issues with out-of-sample prediction, we expect it might work well even for a slightly misspecified model.

Although it does well at recovering known parameters, many readers may be more concerned that the network-based estimation process lines up with other methods of estimating the model, such as the hierarchical Bayesian approach we used above. To test this, we fit the same data from the predictive validity studies above – the Race IAT and Sexuality IAT from (Buttrick et al., 2020) – using the neural network. We then compared the resulting estimates against those from the hierarchical Bayesian approach.

The results are shown in Fig. 11 as correlations between standardized (z-scored) parameter values. Because the scale of a model is fixed by the value of the diffusion rate and the step size used to approximate the diffusion process, the parameters will not necessarily be on the same scale – hence the standardization. In general, the two modeling approaches lined up fairly well, with all correlations around .5 or higher. Note that both methods are good at estimating the true parameters, in that they recover known values from simulations. However, they both have imperfect correlations with the true values, meaning that both the predictors (e.g., Bayesian estimates) and predicted values (e.g., neural network estimates) shown in Fig. 11 have noise. As a result, the correlation between models is lower than the correlation between a model and the true parameter values. Specifically, there is a natural upper limit to how well they can correspond to one another, which is how well they can correspond to themselves (i.e., the reliability). Therefore, we computed both the linear correlation r between network-based and Bayesian (MCMC)-based fitting approaches as well as the corrected correlation ρ. These are shown in the top-left and the bottom-right of each panel, respectively.

In general, we have found that comparisons between hierarchical Bayesian and neural network-based approaches to model fitting result in relative parity between the two approaches. The major difference is the hierarchical Bayesian approach tends to induce some shrinkage, drawing estimates closer to the group-level mean compared to the neural net-

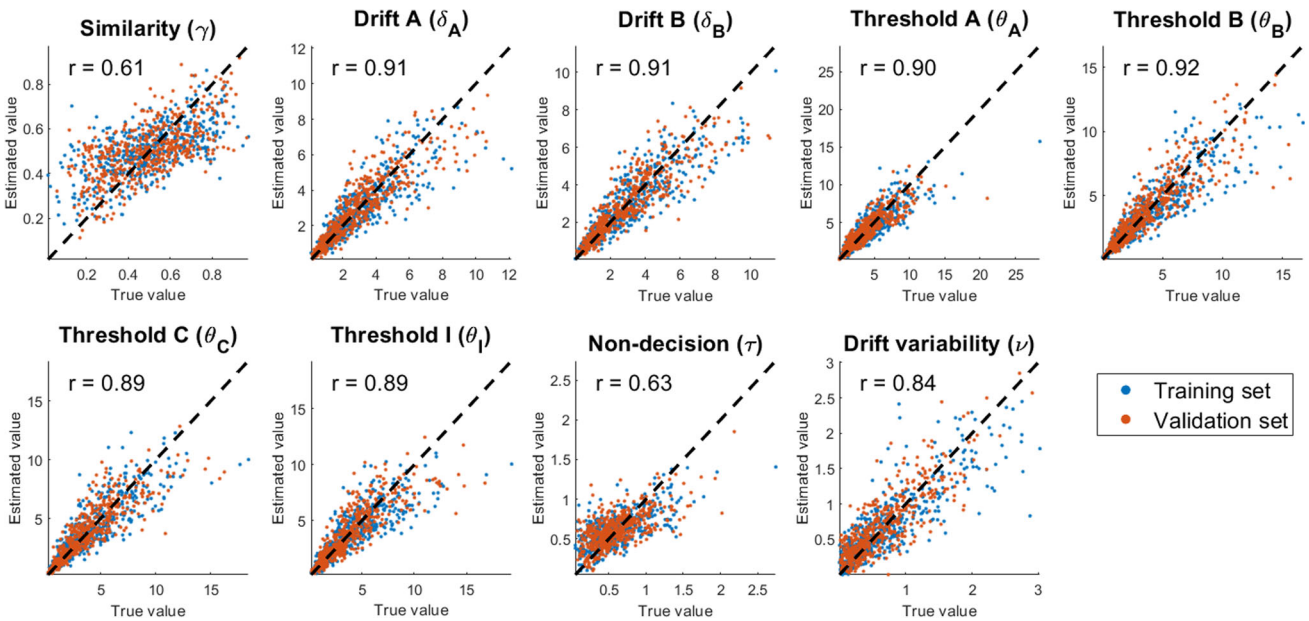

Fig. 10 Relationship between the true parameter values (x) and the estimated parameter values from the neural network (y) for each parameter of the GSR-DDM

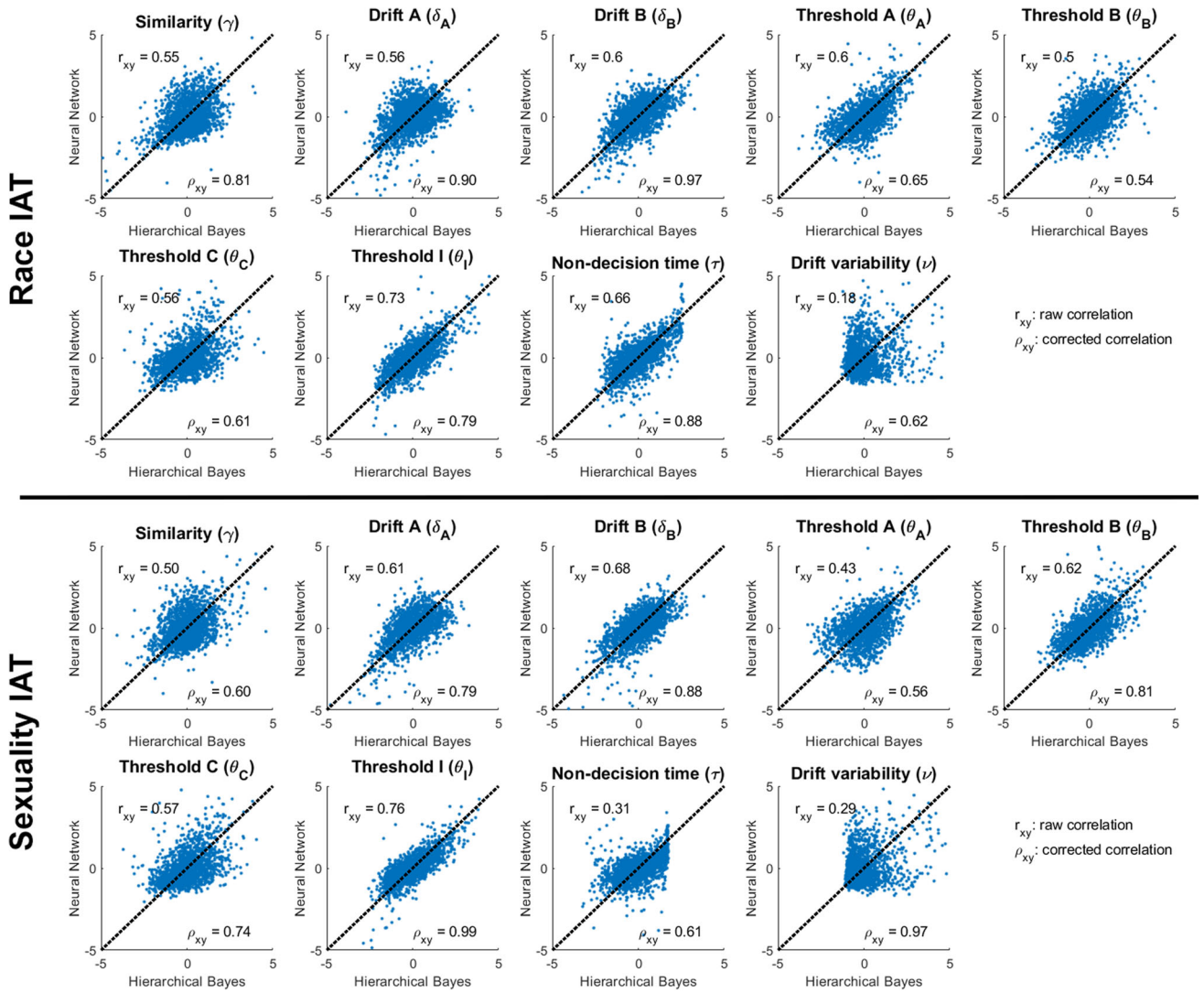

Fig. 11 Comparison between standardized estimates generated from a hierarchical Bayesian fitting method (x) and the new neural network fitting method (y) for each parameter of the GSR-DDM

work estimates. We leave the question of whether this is desirable to the user, but note that it often results in slightly lower correlations between true and estimated parameters (i.e., slightly worse recovery) for the hierarchical Bayesian approach.

Availability

The estimates from the neural network appear to correspond well to both true values (recovery) and estimates from a hierarchical Bayesian fitting approach. However, much of its value is in its ability to immediately fit the data. In terms of fitting time, the hierarchical Bayesian model takes 2-4 hours combined to fit the two Buttrick data sets, which is reasonable given the hierarchical constraints on over 3700 participants. By comparison, the neural network takes about 15 seconds. This speed-up is typical of neural network approaches, which we have found to be on the order of 500-1000 times for large data sets.

Because the trained neural network operates so quickly, it is possible to embed it within online tools for model fitting. Using MATLAB Compiler, we created a downloadable application that allows users to upload their data, estimate the model parameters, visualize the results, and save the resulting parameter estimates for each participant in their data set. This app is available as an executable on the OSF page for this paper at `osf.io/znsfb`. We strongly urge users to follow the instructions in the readme file before using it on their own data, and to be sure to inspect the resulting estimates – and ideally compare them to estimates from a hierarchical Bayesian implementation – to ensure they are sensible before using them in any subsequent analyses.

Our hope is that by making a simple point-and-click web app for model fitting these tools will become more widespread. The use of computational modeling has the potential to make IATs much more effective measurement tasks and, as we show above, shows clear promise in terms of improving their reliability and predictive validity.

Discussion

In this paper, we introduced three main innovations. First, we developed a model of performance on the IAT (GSR-DDM) that put together cutting-edge models of decision-making and similarity representation. This model teased apart conceptual similarity - arguably the construct that the IAT is designed to measure - from processing speeds for different stimuli, control processes related to response caution, and processes like stimulus encoding that each also make contributions to behavior on the IAT. In doing so, it allows for a more in-depth and complete understanding of what participants are doing on IATs, and improves our ability to quantify individual differences in performance in a meaningful way.

Second, we introduced a hierarchical Bayesian method for model fitting and for estimating the test-retest reliability of model parameters. This approach allows us to characterize both individual differences in performance (e.g., similarity representations) alongside group-level trends (e.g., differences in response caution between conditions, general tendencies toward anti-Black mental representations), as well as to estimate the covariance and uncertainty in performance across multiple testing sessions. By virtue of using hierarchical Bayesian estimation, we did not require the large volume of data that other dynamic modeling approaches do (Röhner and Lai, 2021; Klauer et al., 2007). As a result, this model is widely applicable to the deep IAT literature whose foundation is built on traditional paradigms with only 60 total trials for the compatible or incompatible conditions (and only 20 or 40 per testing block within each condition), rather than restricting our inferences to a limited set of specific data sets featuring a large number of trials or heavy time pressure to induce mistakes (Calanchini and Sherman, 2013).

Finally, we developed a neural-network implementation of the model, which sped up the fitting process by nearly 1000 times over and allowed it to be embedded within a freely-available web app. We showed that this approach consistently arrived on the true underlying parameters in a model recovery study, and that it arrived at similar estimates to the hierarchical Bayesian approach on the two large Buttrick et al. (2020) data sets.

Put together, the modeling and estimation approaches allow us to reliably quantify performance on the IAT in terms of the processes we are trying to measure, use these measures to better predict real-world outcomes like contact with people from minoritized social groups, and made it possible to use these approaches without requiring extensive training on computational modeling or coding. As an added bonus, the model conferred insights that would not be possible without the modeling. For example, participants appear to be more careful when completing trials during the incompatible block (higher thresholds), which would produce weaker IAT effects while being undetectable when using purely response time-based measures.

Our results suggest that the "reliability paradox" (Hedge et al., 2018) and parallel criticisms of the IAT as an unreliable method for capturing individual differences (Banse et al., 2001) are largely a matter of measurement. The *D*-score and other simple metrics that attempt to summarize the richness of behavior with a single metric naturally miss many psychologically meaningful aspects of task performance. These metrics, as well as contrast-based measures of individual differences like between-condition parameter differences from the diffusion model, contain a large amount of noise when compared to GSR-DDM parameters, and thus provide a highly impoverished view of individual-level behaviors that go into generating IAT data see Vadillo et al. (2021).

Construct validity and selective influence

One of the major reasons that the new model presented here was able to out-perform other approaches is that it quantifies behavior in terms of more (meaningful) psychological processes. In other words, it appears to have greater *discriminant validity* than other accounts of behavior. To truly test whether the model parameters index separable psychological and cognitive processes, we also examined the correlations among GSR-DDM model parameters. The results are provided in Appendix A. In short, the two drift rates are highly correlated in both the Race and Sexuality IATS, likely corresponding to a domain-general measure of processing speed or intelligence (Lerche et al., 2020; Schubert et al., 2015). This provides convergent validity for the drift rates, as correlations among within-person drifts across tasks and conditions is something we should expect to find in any study (Schubert et al., 2017).

There are otherwise only weak correlations among model parameters, suggesting that they describe distinct components of performance on IATs. In particular, the conceptual-similarity parameter that is central to the GSR-DDM was unrelated to any other parameters. Put together with the finding that conceptual-similarity has the greatest predictive validity of any of the model parameters, it is clear that it is a critical component of the model and an element of performance that ought to be delineated from other contributions to performance like response caution. Overall, the model exhibits a high degree of discriminant validity across its different parameters, reinforcing the proposition that there

are many components of behavior on the IAT that ought to be differentiated in a valid model.

Fortunately, there seems to be sufficient evidence of discriminant validity and its separability from other model parameters, as shown in Appendix A. This evidence is compounded by common-sense observations about most of the model parameters: drift rates (but not conceptual-similarity) increases over time reflecting practice effects, drifts for words vs faces differs in the expected direction, parameter recovery is consistently successful for each of the estimation methods we examined, and the conceptual-similarity parameter is restricted to affect performance only in the paired-response conditions (and in fact is the only parameter that can capture positive covariance between speed and accuracy in these conditions!) yet the model accounts well for performance in these conditions.

It is important to note that the restrictions we placed on the parameters, and which ones change across conditions, is critical to interpreting them. Both base drifts and conceptual-similarity are parameters that ultimately feed into drift rates to create the (statistical) Wiener distribution of response times. This means that it is not possible to allow them both to vary freely across conditions, as a completely unrestricted model would not be identifiable – in the same way that a diffusion model without the within-trial noise (or some other) parameter would not have a fixed scale. Note that this confusion only occurs for binary choice paradigms: these parameters are clearly distinguishable and uniquely identifiable continuous-response or multi-alternative choice, where drift magnitude and drift direction correspond to base drift and conceptual similarity (Kvam et al., 2023). Future work could explore the formation and change of attitudes and conceptual-similarity using paradigms like reinforcement learning, where participants learn an association between stimuli and valence (or two stimulus features) and then take an IAT where performance is quantified using the GSR-DDM. There is some evidence that it is at least theoretically possible to induce secondary associations with unconscious conditioning (Greenwald and De Houwer, 2017). If the experimental paradigm induces a strong enough association, then we should expect it to show up in estimates of γ in such a task. This type of study is a substantial undertaking and well outside the scope of the current paper, but we look forward to future work exploring this possibility.

Implications of GSR-DDM for the IAT's "implicitness"

As noted throughout this paper, the ongoing (and, generally, unexamined) assumption that a participant's responses on the IAT map directly onto something "implicit" has led us into conceptual confusion. It is important to remember that the IAT is a measurement procedure and does not have a one-to-one relationship to the psychological construct of "implicit bias" (Gawronski et al., 2022). Further, it is important for researchers to state in what way they mean that something is implicit (Corneille and Hütter, 2020). Recently, a consensus appears to be building that the conceptual focus of implicitness is more usefully organized around intentionality rather than awareness which, arguably, was the primary focus for the first two decades of research using the IAT see De Houwer and Boddez (2022); Dovidio and Kunst (2022); Krajbich (2022); Melnikoff and Kurdi (2022); Olson and Gill (2022); Ratliff and Smith (2022). Although GSR-DDM does make substantial progress in quantifying conceptual-similarity and its role on IATs, it does not speak fully to the "implicit" (e.g., automatic, unconscious, unintentional) nature of the IAT (Corneille and Hütter, 2020; Gawronski et al., 2022). Traditionally, thresholds (θ) have been thought of as processes that are under the control of the decision maker (Ratcliff et al., 2016), as they describe the level of caution a person takes. Thresholds are thought to vary across blocks, so that participants can set them strategically based on the directions they are given, but not from trial to trial based on the stimuli they see. These thresholds are often described as "sticky" because they cannot be adjusted quickly with learning, directions, or even incentives (Larson and Hawkins, in press; Fontanesi et al., 2019a). A participant would almost certainly have to strategically set their threshold higher in the incompatible condition for the model to detect an effect, given the limited number of trials on most IATs. Whether or not a participant will explicitly admit to having a bias, many participants clearly recognize that the incongruent condition will be more difficult for them and set their thresholds accordingly.

Even without assuming that changes in thresholds are driven by changes in strategy, we can at least say that observing a threshold shift should increase our belief that participants are aware of or control their performance between conditions. Specifically, if we want to make an inference about the likelihood of participants being aware (strategically shifting their thresholds, Pr(strategic)) based on an observed change in the threshold $\Delta\theta$, we can use Bayes rule to update our beliefs:

$$\Pr(\text{strategic} \mid \Delta\theta) = \frac{\Pr(\Delta\theta \mid \text{strategic}) \cdot \Pr(\text{strategic})}{\Pr(\Delta\theta)}$$

. A threshold change should signal awareness by increasing the strength of our beliefs in strategic manipulation from the prior (Pr(strategic)) to the posterior (Pr(strategic $\mid \Delta\theta$)) whenever $\Pr(\Delta\theta \mid \text{strategic}) > \Pr(\Delta\theta)$. In essence, a threshold change indicates strategic manipulation of choice strategy so long as thresholds are more likely to change when participants are strategically manipulating them. Given the deep literature showing that participants successfully strategically manipulate their thresholds under instructions to do

so (Wickelgren, 1977; Pew, 1969; Heitz, 2014; Ratcliff et al., 2016; Heathcote and Matzke, 2022; Donkin and Brown, 2018), this is effectively guaranteed. Detecting a difference in thresholds when participants are deliberately changing their strategies, $Pr(\Delta\theta \mid \text{strategic})$, can be expected to occur much more often than spontaneous differences in thresholds, $Pr(\Delta\theta)$, ultimately providing strong support for strategic manipulation when threshold changes are detected. We can therefore infer, based on the available evidence, that participants exhibiting this effect are likely to be deliberately changing their strategies between congruent and incongruent conditions.

Based on this evidence, the differences we observe between conditions in thresholds (e.g., highest in incompatible blocks) should correspond to biases for which participants have some level of awareness. Previous research estimating similar threshold parameters used the literature on "faking" IAT task performance as their starting point (Röhner and Ewers, 2016). Behavior on the IAT is less controllable than responses to analogous self-report measures (e.g., thermometer ratings), but participants can "fake" their IAT scores, at least under some conditions (Fiedler and Bluemke, 2005; Kim, 2003; Steffens, 2004). The most straightforward method for faking IAT scores is a combination of consciously slowing down during the compatible blocks and speeding up during the incompatible blocks (Cvencek et al., 2010; Fiedler and Bluemke, 2005; Röhner et al., 2013). Accordingly, the response caution parameter from the diffusion model is especially relevant to identifying faking behavior. For example, (Röhner and Ewers, 2016) used a simplified diffusion model and found that participants who were instructed to fake their IAT performance – with or without being given explicit directions as to how – showed greater response caution on the incompatible block trials. Likewise, our findings indicated that the threshold parameter was relevant primarily for the incompatible trials.

Consistent with the idea that response caution was affected by intentions to fake IAT scores, people with less prior contact with minoritized group members had higher thresholds for the incompatible trials on the Race and Sexuality IATs, respectively. Yet, people with greater internal motivation to respond without prejudice actually had a lower incompatible threshold for the Sexuality IAT. The most coherent explanation for this is that people with an intrinsic motivation to respond without prejudice are successful at doing so – that is, they wind up having fewer negative associations with minoritized groups and thus a lower similarity parameter (close to zero), as shown in Table 2. As a result, there is less of a need to "fake" performance on the Sexuality IAT by increasing their thresholds on incongruent trials, and thus lower thresholds for Motivation on these trials compared to other participants who know they might be biased and have difficulty with this condition (Table 2). Conversely, a similar effect did not occur for the Race IAT, indicating that internal motivation to control prejudice may not confer the same advantage across all target social groups. Given that the Race IAT and implicit race bias are the focal point of Project Implicit where these data were collected (Xu et al., 2014), people who are more motivated to control racial prejudice may set relatively higher thresholds because they recognize how important the incompatible pairings are in the context of Project Implicit. Accordingly, they exercise additional response caution that puts them more in line with people with lower motivation, albeit for different reasons.

Although thresholds are typically deliberately controlled, it is less clear whether drift and similarity in the model are driven by features of the stimulus or by an underlying attitude. It may be the case that participants are aware of their own biases that are measured by the association parameter, but unable to control them when elicited via the IAT, or it could be that they simply do not know that they hold these negative associations. For those participants who shifted their threshold in the incongruent condition, we can be more certain that they are aware of the bias on some level. That is, if it is unimportant for a person to appear unprejudiced, then they should have no reason to go through the trouble of implementing a strategy that makes their IAT scores indicate less bias. Of note, the apparent presence of bias, and participants' apparent knowledge of them illustrated by the threshold shifts, does not necessarily mean that participants will be able to use explicit measures to report on the biases that the model captures. This issue of whether one is (or can be made) aware of the mental contents and/or processes indexed by the IAT [e.g.,] Hahn et al. (2014); Gawronski et al. (2022) has been an exceptionally thorny one over the years, in part because the answer may depend on how one defines the terms (Hahn and Goedderz, 2020). Methodical examination of the threshold parameter in GSR-DDM may add productively to this conversation.

Finally, although the positive correlation between IATs and parallel explicit measures of attitudes is well-established (Nosek, 2007), there is mounting evidence that the overlap is more substantial than previously thought when researchers account for measurement error and ensure that explicit measures provide adequate coverage of the attitudinal domain (Blanton et al., 2016; Schimmack, 2021). Therefore, it is especially valuable to develop modeling approaches that can provide additional insight into the automatic versus controlled processes underlying the IAT. In this case, our approach provided unique insights by GSR-DDM's ability to not only isolate clearly distinguishable parameters that should be controllable or uncontrollable, but also to use the parameters to directly predict multiple explicit outcome measures. By doing so, we opened another avenue for generating new hypotheses for future research; specifically, future work should aim to identify the specific factors that cause response

caution to function differently across people and IAT content.

Alternative approaches

We assumed that the effect of similarity is additive with those of the stimulus in terms of determining drift, and can speed up decision-making in cases where the similarities facilitate a particular response in addition to slowing them down when the similarities are incompatible. However, it may be the case that the total information processing capacity (i.e., our overall ability to update our beliefs over time) is restricted to the point where these similarities cannot be processed in parallel to the target categorization task, or where the overall drift is a dilution of target categorization + similarities. This would mean that the influence of similarities would be limited in how they could intrude on the choice that someone is trying to make. In such a case, even positive similarities might not help information processing because attention would be split between the target task and the similarities between stimuli and irrelevant categories. Put differently, it may be that drift for the target decision (good-bad, or Black-White in the Race IAT) shares a fixed capacity with similarity-driven beliefs (good-bad). One way to resolve this question would be to identify clear facilitation effects (Lindsay and Jacoby, 1994; Heathcote et al., 1991), which would directly contradict an interference-only explanation for performance.

Another important consideration is the ability of the model and experimental paradigm to disentangle positive attitudes toward one category from negative attitudes toward the opposing category. Here, we have treated race, introversion-extraversion, and sexuality each as a single dimension with the categories represented as polar opposites (e.g., Black is opposite White in Fig. 1). Given that Black people are not actually the opposite of White people nor straight people the opposite of gay people, the model's setup gives rise to the potential that the model works best for attitude objects that are naturally bipolar (e.g., Democrats vs. Republicans) or for people who have a tendency to see groups as being opposite from one another. For example, we might posit better model fit for those high in essentialism (Haslam et al., 2000; Prentice and Miller, 2007). Interestingly, IATs for bipolar pairs of attitude objects correspond more closely to self-reported attitudes than IATs for more unipolar pairs (Nosek, 2005).

IATs allow us to make inferences about the relative valence of the two categories by comparing compatible and incompatible conditions. We could weaken the overall effect by increasing similarity with the minoritized category, or by making more dissimilarities with the majority category (Dasgupta and Greenwald, 2001; Joy-Gaba and Nosek, 2010). However, the base IAT does not allow us to examine the association of one category with positive / negative in absence of the other category, and thus we cannot estimate separate similarities for each of the target categories. However, the model could in principle incorporate multiple similarities, including (e.g., for the race IAT) White-positive, White-negative, Black-positive, Black-negative, and even White-Black. The cognitive representations of these categories are high-dimensional, and are formed by many examples and pairings that people encounter across their lives (Deerwester et al., 1990; Kvam, 2019a). To get at each of the separate similarities, we would need richer behavioral data. This could involve looking at decisions consisting of only three of the options, as in the single target IAT (Bluemke and Friese, 2008), comparing multiple single-target IAT conditions like those for neutral / reference stimuli, adopting approaches like the word-embedding association test (Caliskan et al., 2017) to inform our estimates of multiple similarity parameters, or even doing latent semantic analysis on a large corpora of text all collected from one person (Landauer and Dumais, 1997; Landauer, 2006) to estimate these similarities a priori. Certainly, it would be an interesting challenge to relate the model parameters to independent assessments of similarity.

Another potential extension of our approach would be to put it together with multi-stage models of decision making, accounting for dual or sequential influences of automatic and controlled processes. This could also bring together the disparate lines of work looking at multinomial processing trees, which are inherently multi-stage (Calanchini and Sherman, 2013; Meissner and Rothermund, 2013), with the type of single-process, response-time focused models like the one we present here. This might be accomplished by having evidence accumulation unfold in multiple stages as in work by Diederich and Trueblood (2018), or could extend MPTs with response time models embedded into each branch as in work by Klauer and Kellen (2018). Increasing start point variability or having a contaminant guessing process would also bring the model more closely in line with models like the Quad model (Calanchini and Sherman, 2013); however, the lack of fast errors in the IATs we analyzed suggest that guessing may not play a significant role, at least in this data set (Ratcliff, 1985; Ratcliff et al., 2016).

Individual differences in parameter estimates

The relationship between individual differences and estimated model parameters has not been tested in any meaningful way and could make an important theoretical contribution on its own. In one example, using a similar modeling approach as the one we propose, those higher in Need for Closure (Roets and Van Hiel, 2007) have a lower decision threshold in a speed-accuracy tradeoff task when speed was emphasized, but not when accuracy was emphasized (Evans

et al., 2017).[5] As such, testing the relationship between Need for Closure and the threshold parameter in the current model may prove be useful.

Closer to the current work, Calanchini et al. (2014) speculate that their observation that the Quad model's Detection parameter is correlated across attitude domains may indicate that it is related to individual differences in motivation or ability to focus on the task. They further suggest that the overlap in AC (association activation) parameters, even for highly-unrelated IATs, may be due to those higher in Need to Evaluate (Jarvis and Petty, 1996) having stronger activation of associations across all tasks. Both of these ideas remain untested, but they highlight how a focus on modeling processes that underlie behavior on an IAT can lead to new hypotheses as compared to when a summary score such as the IAT *D*-score is the unit of analysis. Given the lack of previous theorizing in this vein, the following is highly speculative. That said, it is worth testing the prediction of Calanchini et al. (2014) that Need for Evaluation correlates positively with GSR-DDM's similarity parameter. In addition, it seems likely that the association / similarity parameter would be stronger for those higher in Need for Affect (Maio and Esses, 2001) due to the relationship between affect and the IAT [e.g.,] Smith and Nosek (2011), as well as for those whose behavior related to the measured construct is habitual (Verplanken and Orbell, 2003). Any variable that increases the accessibility of attitudes [see] Fazio (1995) would be likely to increase the utility of the similarity parameter in GSR-DDM, as it changes the decision space in which the evidence accumulation process unfolds (Kvam, 2019a). We can make additional predictions for the threshold parameter such as that it will be lower for those who are more likely to make decisions based on a reliance on intuition (Pacini and Epstein, 1999) or spontaneity (Scott and Bruce, 1995). Again, these predictions are speculative, but highlight a central benefit of the described modeling approach in that it allows for tests of this type of theoretically-informed speculation.

Limitations and future directions

Although evidence accumulation processes for modeling are now ubiquitous within decision-making (Busemeyer et al., 2019; Ratcliff et al., 2016), their adoption in social psychological tasks has only just begun (Johnson et al., 2017; Pleskac et al., 2018; Röhner and Lai, 2021). As further work is carried out on specific tasks, our theories of the cognitive and social processes that are involved will naturally improve. While the GSR-DDM provides a step forward – an accessible and effective step, we hope – we are certain it will not be the last. In particular, the similarity parameter serves only as a first-pass method at quantifying the role of representational overlap in IAT performance. There are undoubtedly multiple similarities at play, and deeper ways to quantify representational similarity using neural measures (Kriegeskorte et al., 2008).

The neural network fitting approach we used (Radev et al., 2020; Sokratous et al., 2022) is also in its infancy, having only been developed over the past few years. While it already shows impressive performance compared to traditional modeling methods – yielding equally-precise estimates in a tiny fraction of the time – it is best used as a method for simulation-based models that are frequently applied to a common task. This makes it an excellent fit for modeling IATs, but there is undoubtedly room for improvement in terms of optimizing the structure of the neural network (number of nodes, layers, etc.), estimating the error in model parameters (Radev et al., 2020a), and comparing different models (Radev et al., 2021). We are hopeful that more modelers will adopt this approach in order to make modeling as a whole more accessible.

One element of dynamic decision models that we did not approach in this paper is starting point biases. Both multinomial processing trees, like the Quad model, and dynamic models often include a parameter that quantifies a participant's general tendency or bias to respond on the left or right side. This can occur when there is a general bias toward responding 'positive' on incongruent trials – on the premise that it is worse to accidentally categorize a black face as positive than it is to accidentally categorize it as negative. Empirically, there did not appear to be an overall bias toward one side or another in the data sets that we used, which is why we did not include this parameter in the current model. However, there are certainly data where this would be useful to include, and potentially IATs where this would be important to control and account for. Certainly paradigms where there are manipulations of base rates, incentives for different responses (including those conferred by social interactions), or predecision information provided would constitute cases where start point biases would be important to consider (Diederich and Busemeyer, 2006; Axt and Johnson, 2021; Heathcote et al., 2019).

Conclusion

In sum, our newly-developed GSR-DDM model describes behavior on IATs in a more complete and accurate way not only compared to simple summary metrics like IAT *D*-scores but also other previous attempts to model the distinct processes underlying IAT task performance. We gained unique insights into test-retest reliability and predictive validity both for the similarity parameter that is most reflective of the

[5] It is notable that instructions presented before the IAT commonly instruct participants with some version of "go as fast as you can, while making as few mistakes as possible", which leaves the relative importance of speed and accuracy open for the participant.

relatively automatic associative processes that researchers typically intend to capture with the IAT, as well as processes that may or may not be subject to participants' control or awareness. Consistent with the most common theoretical accounts of the IAT (Greenwald and Banaji, 2017), associations or conceptual similarities are a central part of GSR-DDM and were most important for predicting relevant outcomes [but see] De Houwer et al. (2021). These similarities are formed and revised over time through both classical and operant learning processes, they correspond to meaningful individual differences, and they affect choices that people make in the real world. The new approaches we developed here should make them more accessible to researchers working on these problems, improving the utility of IATs and the ease of modeling the cognitive processes involved in performance on these tasks.

Funding This work was supported by a SEED grant from the University of Florida Informatics Institute (UFII) to PDK, a UFII graduate fellowship to KS, and National Science Foundation grants to CTS (BCS-2125944) and PDK (SES-2237119).

Code Availability The model and analysis code for this project are freely available on the Open Science Framework at `osf.io/znsfb/`.

Declarations

Ethics Approval This research used fully de-identified secondary data and was therefore not considered human subjects research.

Consent to Participate Not applicable.

Consent for Publication The authors consent to publication of this paper and assert that it is not currently under consideration elsewhere.

Conflicts of interest CTS serves as a member-at-large on the Scientific Advisory Board of Project Implicit, Inc. No other authors declare conflicts of interest related to this work.

Appendix A: Correlations among model parameters

One concern among models in general is their ability to index separable psychological and cognitive processes. To do so, we can examine how strongly the estimates of different parameters are related to one another. Theoretically, parameters that index common or related cognitive processes will have high correlations with one another while ones that are unrelated or easily malleable are more likely to have low correlations with one another. Therefore, it is useful to look at the inter-parameter correlations when assessing the discriminant and convergent validity of the model.

The correlations among model parameters for the GSR in the Race IAT (Buttrick et al., 2020) are shown in Table 4. The highest correlation is between the two drift rates, at $r = .69$. This is perfectly in line with what we should expect, as correlations among drift rates reflect a general tendency toward overall faster or slower information processing across participants (Lerche et al., 2020). Aside from this, most correlations among parameters are weak, with some stronger ones between thresholds or drift rates and the drift variability parameter ν. Again, this is a relatively common finding in diffusion models, and not one that should give us much pause. The lower discriminant validity is part of why we do not focus much on the drift rate variability parameter, as much like non-decision time it tends to be a "nuisance" parameter that is included to account for extraneous factors in order to improve the overall fit of the model and recovery of other parameters (Steingroever et al., 2020).

Table 3 Correlations among GSR-DDM parameters for the Race IAT. These parameters are conceptual similarity γ, drift rates for faces δ_F, drift rates for words δ_W, threshold for words-only condition δ_{+-}, threshold for faces-only condition δ_{BW}, threshold for black-negative/white-positive condition θ_{B-}, threshold for black-positive/white-negative condition θ_{B+}, non-decision time τ, and drift variability ν

	γ	δ_F	δ_W	θ_{+-}	θ_{BW}	θ_{B-}	θ_{B+}	τ	ν
γ	1								
δ_A	0.07	1							
δ_B	0.07	0.69	1						
θ_{+-}	-0.13	-0.41	-0.3	1					
θ_{BW}	-0.07	-0.32	-0.28	0.26	1				
θ_{B-}	0.29	-0.21	-0.27	0.09	0.13	1			
θ_{B+}	-0.22	-0.38	-0.39	0.14	0.07	0.07	1		
τ	0.09	0.12	0.09	0.05	0.14	0.24	0.21	1	
ν	-0.01	-0.38	-0.31	0.33	0.43	0.4	0.29	0.22	1

Finally, there are some weak correlations between the thresholds in "basic" IAT conditions (valence-only "+-" condition, and black-white "BW" condition) and drift rates for the different types of stimuli (δ_F for faces and δ_W for words). This can reflect either participants' informed expectations about their own performance – participants who know they will perform well (higher drifts) on a basic task like these conditions are able to set lower thresholds while maintaining a high level of performance – or correlations among parameters that sometimes occur with differences in speed-accuracy manipulations (Donkin et al., 2014). All other correlations were fairly weak and within acceptable limits for discriminant validity in cognitive modeling (Heathcote et al., 2015).

The correlations among model parameters for the GSR in the Sexuality IAT (Buttrick et al., 2020) are shown in Table 4. The findings are almost exactly the same, lending credibility to the conclusions we drew from the Race IAT. As in the Race IAT, the highest correlation among model parameters is between the two drift rates, at $r = .67$, reflecting a general tendency toward overall faster or slower information process-

Table 4 Correlations among GSR-DDM parameters for the Sexuality IAT. These parameters are conceptual similarity γ, drift rates for sexuality words δ_A, drift rates for valence words δ_B, threshold for valence-only condition δ_{+-}, threshold for sexuality-only condition δ_{GS} (G = gay, S = straight), threshold for gay-negative/straight-positive condition θ_{G-}, threshold for gay-positive/straight-negative condition θ_{G+}, non-decision time τ, and drift variability ν

	γ	δ_A	δ_B	θ_{+-}	θ_{GS}	θ_{G-}	θ_{G+}	τ	ν
γ	1								
δ_A	0.04	1							
δ_B	0.01	0.67	1						
θ_{+-}	-0.08	-0.37	-0.28	1					
θ_{GS}	-0.11	-0.35	-0.31	0.29	1				
θ_{G-}	0.07	-0.25	-0.18	0.12	0.25	1			
θ_{G+}	-0.11	-0.28	-0.2	0.11	0.25	0.09	1		
τ	0.02	0.11	0.25	0.05	0.19	0.22	0.28	1	
ν	-0.09	-0.42	-0.3	0.39	0.51	0.45	0.38	0.23	1

ing on the task. Also as before, there were some correlations with drift rate variability and some between threshold and drift rates that should not be too concerning (Steingroever et al., 2020). All other correlations were fairly weak and within acceptable limits for discriminant validity in cognitive modeling.

Perhaps most importantly and most notably, the correlations between the conceptual-similarity parameter γ and all other parameters, in both IATS, was extremely low. This suggests that the most central parameter of our model was distinguishable from all other parameters, meaning it indexes a unique component of performance on IATs that is not clearly captured by the classic DDM. These results therefore emphasize the discriminant and convergent validity of our model – parameters that are meant to be distinct (e.g., γ) are uncorrelated with others, while those that are clearly related (e.g., drift rates) are tightly correlated.

References

Axt, J. R., & Johnson, D. J. (2021). Understanding mechanisms behind discrimination using diffusion decision modeling. *Journal of Experimental Social Psychology, 95*, 104134.

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of personality and social psychology, 97*, 533.

Bading, K., Stahl, C., & Rothermund, K. (2020). Why a standard iat effect cannot provide evidence for association formation: The role of similarity construction. *Cognition and Emotion, 34*, 128–143.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the iat. *Zeitschrift für experimentelle Psychologie, 48*, 145–160.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of cognitive neuroscience, 8*, 551–565.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. *Problems in measuring change, 2*, 3–20.

Bhatia, S. (2013). Associations and the accumulation of preference. *Psychological Review, 120*, 522–543.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review, 124*, 1–20.

Bhatia, S., & Mullett, T. L. (2018). Similarity and decision time in preferential choice. *Quarterly Journal of Experimental Psychology, 71*, 1276–1280.

Blanton, H., Burrows, C. N., & Jaccard, J. (2016). To accurately estimate implicit influences on health behavior, accurately estimate explicit influences. *Health psychology, 35*, 856.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42*, 192–212.

Bluemke, M., & Friese, M. (2008). Reliability and validity of the single-target iat (st-iat): assessing automatic affect towards multiple attitude objects. *European journal of social psychology, 38*, 977–997.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877–1901.

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in Cognitive Sciences, 23*, 251–263.

Buttrick, N., Axt, J., Ebersole, C. R., & Huband, J. (2020). Re-assessing the incremental predictive validity of implicit association tests. *Journal of Experimental Social Psychology, 88*, 103941.

Cai, H., Sriram, N., Greenwald, A. G., & McFarland, S. G. (2004). The implicit association test's d measure can minimize a cognitive skill confound: Comment on mcfarland and crouch (2002). *Social Cognition, 22*, 673–684.

Calanchini, J., Meissner, F., & Klauer, K. C. (2021). The role of recoding in implicit social cognition: Investigating the scope and interpretation of the real model for the implicit association test. *PloS one, 16*, e0250068.

Calanchini, J., & Sherman, J. W. (2013). Implicit attitudes reflect associative, non-associative, and non-attitudinal processes. *Social and Personality Psychology Compass, 7*, 654–667.

Calanchini, J., Sherman, J. W., Klauer, K. C., & Lai, C. K. (2014). Attitudinal and non-attitudinal components of iat performance. *Personality and Social Psychology Bulletin, 40*, 1285–1296.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*, 183–186.

Carlsson, R., & Agerström, J. (2016). A closer look at the discrimination outcomes in the iat literature. *Scandinavian journal of psychology, 57*, 278–287.

Carpenter, T. P., Goedderz, A., & Lai, C. K. (2022). Individual differences in implicit bias can be measured reliably by administering the same implicit association test multiple times. Personality and Social Psychology Bulletin, (p. 01461672221099372).

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature neuroscience, 14*, 1462–1467.

Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science, 32*, 218–240.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: the quad model of implicit task performance. *Journal of personality and social psychology, 89*, 469.

Corneille, O., & Hütter, M. (2020). Implicit? what do you mean? a comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review, 24*, 212–232.

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences, 117*, 30055–30062.

Cvencek, D., Greenwald, A. G., Brown, A. S., Gray, N. S., & Snowden, R. J. (2010). Faking of the implicit association test is statistically detectable and partly correctable. *Basic and applied social psychology, 32*, 302–314.

Cybenko, G. V. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems, 2*, 303–314.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of personality and social psychology, 81*, 800.

De Houwer, J. (2006). What are implicit measures and why are we using them. The handbook of implicit cognition and addiction, (pp. 11–28).

De Houwer, J., & Boddez, Y. (2022). Bias in implicit measures as instances of biased behavior under suboptimal conditions in the laboratory. *Psychological Inquiry, 33*, 173–176.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological bulletin, 135*, 347.

De Houwer, J., Van Dessel, P., & Moran, T. (2021). Attitudes as propositional representations. *Trends in Cognitive Sciences, 25*, 870–882.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science, 41*, 391.

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2020). The case for formal methodology in scientific reform. *Royal Society open science, 8*, 200805.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: the role of motivations to respond without prejudice. *Journal of personality and social psychology, 82*, 835.

Diederich, A., & Busemeyer, J. R. (2006). Modeling the effects of payoff on response bias in a perceptual discrimination task: Bound-change, drift-rate-change, or two-stage-processing hypothesis. *Perception & Psychophysics, 68*, 194–207.

Diederich, A., & Trueblood, J. S. (2018). A dynamic dual process model of risky decision making. *Psychological Review, 125*, 270–292.

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E.-J. (2011). Diffusion versus linear ballistic accumulation: Different models but the same conclusions about psychological processes? *Psychonomic Bulletin & Review, 18*, 61–69.

Donkin, C., & Brown, S. D. (2018). Response times and decision-making. *Stevens' handbook of experimental psychology and cognitive neuroscience, 5*, 349–377.

Donkin, C., Little, D. R., & Houpt, J. W. (2014). Assessing the speed-accuracy trade-off effect on the capacity of information processing. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 1183.

Dovidio, J. F., & Kunst, J. R. (2022). Delight in disorder: Inclusively defining and operationalizing implicit bias. *Psychological Inquiry, 33*, 177–180.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2016). The development of implicit gender attitudes. *Developmental Science, 19*, 781–789.

Evans, N. J., Rae, B., Bushmakin, M., Rubin, M., & Brown, S. D. (2017). Need for closure is associated with urgency in perceptual decision-making. *Memory & cognition, 45*, 1193–1205.

Fazio, R. H. (1995). Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. *Attitude strength: Antecedents and consequences, 4*, 247–282.

Fengler, A., Govindarajan, L. N., & Frank, M. J. (2020). Encoder-decoder neural architectures for fast amortized inference of cognitive process models. cognitivesciencesociety.org.

Fiedler, K., & Bluemke, M. (2005). Faking the iat: Aided and unaided response control on the implicit association tests. *Basic and Applied Social Psychology, 27*, 307–316.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "i", the "a", and the "t" : A logical and psychometric critique of the implicit association test (iat). *European review of social psychology, 17*, 74–147.

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic bulletin & review, 26*, 1099–1121.

Fontanesi, L., Palminteri, S., & Lebreton, M. (2019). Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using diffusion decision modeling. *Cognitive, Affective, & Behavioral Neuroscience, 19*, 490–502.

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., & Lochbaum, K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. In Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 465–480). ACM.

Gardner, R., & Neufeld, R. W. (1987). Use of the simple change score in correlational analyses'. *Educational and Psychological Measurement, 47*, 849–864.

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14*, 574–595.

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022). Implicit bias ≠ bias on implicit measures. *Psychological Inquiry, 33*, 139–155.

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin, 43*, 300–312.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science, 7*, 457–472.

Goldberg, Y., & Levy, O. (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. arXiv preprint http://arxiv.org/abs/1402.3722arXiv:1402.3722, .

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *Psychological Bulletin, 75*, 424–429.

Greenwald, A. G. (2017). An ai stereotype catcher. *Science, 356*, 133–134.

Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist, 72*, 861.

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., Hahn, A., Hehman, E., Hofmann, W., & Hughes, S. (2021). Best research practices for using the implicit association test. Behavior research methods, (pp. 1–20).

Greenwald, A. G., & De Houwer, J. (2017). Unconscious conditioning: Demonstration of existence and difference from conscious conditioning. *Journal of Experimental Psychology: General, 146*, 1705.

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*, 1464–1480.

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, 197.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: Iii.

meta-analysis of predictive validity. Journal of personality and social psychology, 97, 17.
Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science, 16*, 789–802.
Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*, 1006–1033.
Gutmann, M. U., & Corander, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research, 17*, 1–47.
Hahn, A., & Goedderz, A. (2020). Trait-unconsciousness, state-unconsciousness, preconsciousness, and social miscalibration in the context of implicit evaluation. *Social Cognition, 38*, s115–s134.
Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*, 1369.
Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. (2021). Learning from the reliability paradox: How theoretically informed generative models can advance the social, behavioral, and brain sciences. PsyArXiv, psyarxiv.com/xr7y3.
Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of social psychology, 39*, 113–127.
Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In An Introduction to Model-Based Cognitive Neuroscience (pp. 25–48). Springer.
Heathcote, A., Holloway, E., & Sauer, J. (2019). Confidence and varieties of bias. *Journal of Mathematical Psychology, 90*, 31–46.
Heathcote, A., & Matzke, D. (2022). Winner takes all! what are race models, and why and how should psychologists use them? *Current Directions in Psychological Science, 31*, 383–394.
Heathcote, A., Popiel, S. J., & Mewhort, D. (1991). Analysis of response time distributions: An example using the stroop task. *Psychological bulletin, 109*, 340.
Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods, 50*, 1166–1186.
Heider, B., & Groner, R. (1997). Backward masking of words and faces: Evidence for different processing speeds in the hemispheres? *Neuropsychologia, 35*, 1113–1120.
Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8*, 150.
Holmes, W. R. (2015). A practical guide to the probability density approximation (pda) with improved implementation and error characterization. *Journal of Mathematical Psychology, 68*, 13–24.
Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology, 27*, 116–159.
Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology, 108*, 187.
Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of personality and social psychology, 70*, 172.
Jin, W., Zhao, L., Zhang, S., Liu, Y., Tang, J., & Shah, N. (2021). Graph condensation for graph neural networks. arXiv preprint http://arxiv.org/abs/2110.07580arXiv:2110.07580, .
Jin, Z. (2016). Disentangling recoding processes and evaluative associations in a gender attitude implicit association test among adult males. *Quarterly Journal of Experimental Psychology, 69*, 2276–2284.
Johns, M., Cullum, J., Smith, T., & Freng, S. (2008). Internal motivation to respond without prejudice and automatic egalitarian goal activation. *Journal of Experimental Social Psychology, 44*, 1514–1519.
Johnson, D. J., Hopwood, C. J., Cesario, J., & Pleskac, T. J. (2017). Advancing research on cognitive processes in social and personality psychology: A hierarchical drift diffusion model primer. *Social Psychological and Personality Science, 8*, 413–423.
Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology, 41*, 137–146.
Kim, D.-Y. (2003). Voluntary controllability of the implicit association test (iat). Social Psychology Quarterly, (pp. 83–96).
Klauer, K. C., & Kellen, D. (2018). RT-MPTs: Process models for response-time distributions based on multinomial processing trees with applications to recognition memory. *Journal of Mathematical Psychology, 82*, 111–130.
Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the implicit association test: Why flexible people have small iat effects. *Quarterly Journal of Experimental Psychology, 63*, 595–619.
Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the implicit association test: A diffusion-model analysis. *Journal of Personality and Social Psychology, 93*, 353–368.
Klein, C. (2020). Confidence intervals on implicit association test scores are really rather large. PsyArXiv, .
Koranyi, N., & Meissner, F. (2015). Handing over the reins: Neutralizing negative attitudes toward dependence in response to reciprocal romantic liking. *Social Psychological and Personality Science, 6*, 685–691.
Krajbich, I. (2022). Decomposing implicit bias. *Psychological Inquiry, 33*, 181–184.
von Krause, M., Radev, S. T., Voss, A., Quintus, M., Egloff, B., & Wrzus, C. (2021). Stability and change in diffusion model parameters over two years. *Journal of Intelligence, 9*, 26.
Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience, (p. 4).
Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and STAN*. Academic Press.
Kruschke, J. K., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review, 25*, 178–206.
Kurdi, B., & Banaji, M. R. (2017). Reports of the death of the individual difference approach to implicit social cognition may be greatly exaggerated: A commentary on payne, vuletich, and lundberg. *Psychological Inquiry, 28*, 281–287.
Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *American psychologist, 74*, 569.
Kvam, P. D. (2019). A geometric framework for modeling dynamic decisions among arbitrarily many alternatives. *Journal of Mathematical Psychology, 91*, 14–37.
Kvam, P. D. (2019). Modeling accuracy, response time, and bias in continuous orientation judgments. *Journal of Experimental Psychology: Human Perception and Performance, 45*, 301–318.
Kvam, P. D., Alaukik, A., Mims, C. E., Martemyanova, A., & Baldwin, M. (2022). Rational inference strategies and the genesis of polarization and extremism. *Scientific reports, 12*, 1–13.
Kvam, P. D., & Busemeyer, J. R. (2020). A distributional and dynamic theory of pricing and preference. *Psychological Review, 127*, 1053–1078.

Kvam, P. D., Marley, A., & Heathcote, A. (2023). A unified theory of discrete and continuous responding. *Psychological Review, 130*, 368–400.

Kvam, P. D., & Pleskac, T. J. (2016). Strength and weight: The determinants of choice and confidence. *Cognition, 152*, 170–380.

Kvam, P. D., Romeu, R. J., Turner, B. M., Vassileva, J., & Busemeyer, J. R. (n.d.). Testing the factor structure underlying behavior using joint cognitive models: Impulsivity in delay discounting and cambridge gambling tasks. Psychological Methods, 26, 18–37.

Kvam, P. D., & Turner, B. M. (2021). Reconciling similarity across models of continuous selections. *Psychological Review, 128*, 766–786.

LaCosse, J., & Plant, E. A. (2020). Internal motivation to respond without prejudice fosters respectful responses in interracial interactions. *Journal of Personality and Social Psychology, 119*, 1037.

Landauer, T. K. (2006). Latent semantic analysis. Wiley Online Library.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review, 104*, 211–240.

Larson, J., & Hawkins, G. (in press). Speed-accuracy tradeoffs in decision making: Perception shifts and goal activation bias decision thresholds. Journal of Experimental Psychology: Lerning, Memory, and Cognition, .

Lemm, K. M. (2006). Positive associations among interpersonal contact, motivation, and implicit and explicit attitudes toward gay men. *Journal of Homosexuality, 51*, 79–99.

Lerche, V., von Krause, M., Voss, A., Frischkorn, G. T., Schubert, A.-L., & Hagemann, D. (2020). Diffusion modeling and intelligence: Drift rates show both domain-general and domain-specific relations with intelligence. *Journal of Experimental Psychology: General, 149*, 2207.

Lin, D. (1998). An information-theoretic definition of similarity. In ICML (pp. 296–304). volume 98.

Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process dissociations: the relationship between facilitation and interference. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 219.

Lindsey, A., King, E., Hebl, M., & Levine, N. (2015). The impact of method, motivation, and empathy on diversity training effectiveness. *Journal of Business and Psychology, 30*, 605–617.

Luce, R. D. (1986). Response Times. 8. Oxford University Press.

Lueckmann, J.-M., Bassetto, G., Karaletsos, T., & Macke, J. H. (2019). Likelihood-free inference with emulator networks.

MacInnis, C. C., Page-Gould, E., & Hodson, G. (2017). Multilevel intergroup contact and antigay prejudice (explicit and implicit) evidence of contextual contact benefits in a less visible group domain. *Social Psychological and Personality Science, 8*, 243–251.

MacLeod, C. M. (1991). Half a century of research on the stroop effect: an integrative review. *Psychological Bulletin, 109*, 163–203.

MacLeod, C. M. (1992). The stroop task: The "gold standard" of attentional measures. *Journal of Experimental Psychology: General 121*, 12–14.

Maio, G. R., & Esses, V. M. (2001). The need for affect: Individual differences in the motivation to approach or avoid emotions. *Journal of personality, 69*, 583–614.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends in cognitive sciences, 7*, 293–299.

Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the implicit association test: the real model for the iat. *Journal of Personality and Social Psychology, 104*, 45.

Meissner, F., & Rothermund, K. (2015). A thousand words are worth more than a picture? the effects of stimulus modality on the implicit association test. *Social Psychological and Personality Science, 6*, 740–748.

Melnikoff, D. E., & Kurdi, B. (2022). What implicit measures of bias can do. *Psychological Inquiry, 33*, 185–192.

Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the implicit association test. *Journal of personality and social psychology, 85*, 1180.

Moors, A., & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. *Psychological bulletin, 132*, 297.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134*, 565.

Nosek, B. A. (2007). Implicit-explicit relations. *Current directions in psychological science, 16*, 65–69.

van Nunspeet, F., Derks, B., Ellemers, N., & Nieuwenhuis, S. (2015). Moral impression management: Evaluation by an in-group member during a moral iat affects perceptual attention and conflict and response monitoring. *Social Psychological and Personality Science, 6*, 183–192.

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the implicit association test: personalizing the iat. *Journal of personality and social psychology, 86*, 653.

Olson, M. A., & Gill, L. J. (2022). Commentary on gawronski, ledgerwood, and eastwick, implicit bias≠ bias on implicit measures. *Psychological Inquiry, 33*, 199–202.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of iat criterion studies. *Journal of personality and social psychology, 105*, 171.

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82*, 85.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of personality and social psychology, 76*, 972.

Paige, K. J., Weigard, A., & Colder, C. R. (2022). Reciprocal associations between implicit attitudes and drinking in emerging adulthood. Alcoholism: clinical and experimental research, 46, 277–288.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry, 28*, 233–248.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology, 90*, 751.

Pew, R. W. (1969). The speed-accuracy operating characteristic. *Acta Psychologica, 30*, 16–26.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of personality and social psychology, 75*, 811.

Pleskac, T. J., Cesario, J., & Johnson, D. J. (2018). How race affects evidence accumulation during the decision to shoot. *Psychonomic Bulletin & Review, 25*, 1301–1330.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing (p. 10). Vienna, Austria. volume 124.

Polich, J., & Donchin, E. (1988). P300 and the word frequency effect. *Electroencephalography and clinical neurophysiology, 70*, 33–45.

Prentice, D. A., & Miller, D. T. (2007). Psychological essentialism of human categories. Current Directions in Psychological Science, (pp. 202–206).

Radev, S. T., D'Alessandro, M., Mertens, U. K., Voss, A., Köthe, U., & Bürkner, P.-C. (2021). Amortized bayesian model comparison with evidential deep learning.

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., & Köthe, U. (2020a). Bayesflow: Learning complex stochastic models with invertible neural networks.

Radev, S. T., Mertens, U. K., Voss, A., & Köthe, U. (2020). Towards end-to-end likelihood-free inference with convolutional neural networks. *British Journal of Mathematical and Statistical Psychology, 73*, 23–43.

Rae, J. R., Reimer, N. K., Calanchini, J., Lai, C. K., Rivers, A. M., Dasgupta, N., Hewstone, M., & Schmid, K. (2020). Intergroup contact and implicit racial attitudes: Contact is related to less activation of biased evaluations but is unrelated to bias inhibition. PsyArXiv, .

Ratcliff, J. J., Lassiter, G. D., Markman, K. D., & Snyder, C. J. (2006). Gender differences in attitudes toward gay men and lesbians: The role of motivation to respond without prejudice. *Personality and Social Psychology Bulletin, 32*, 1325–1338.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review, 92*, 212–225.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*, 873–922.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*, 333–367.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20*, 260–281.

Ratliff, K., & Smith, C. (2021). *Lessons from two decades with project implicit*. A Handbook of Research on Implicit Bias and Racism: APA Books.

Ratliff, K. A., & Smith, C. T. (2022). Implicit bias as automatic behavior. *Psychological Inquiry, 33*, 213–218.

van Ravenzwaaij, D., Brown, S., & Wagenmakers, E.-J. (2011). An integrated perspective on the relation between response speed and intelligence. *Cognition, 119*, 381–393.

van Ravenzwaaij, D., van der Maas, H., & Wagenmakers, E. (2011). Does the name-race implicit association test measure racial prejudice? *Experimental Psychology, 58*, 271–277.

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science, 181*, 574–576.

Reynolds, A., Garton, R., Kvam, P. D., Griffin, V., Sauer, J., Osth, A., & Heathcote, A. (2021). A dynamic model of deciding not to choose. *Journal of Experimental Psychology: General, 150*, 42–66.

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological bulletin, 140*, 1281.

Roets, A., & Van Hiel, A. (2007). Separating ability from need: Clarifying the dimensional structure of the need for closure scale. *Personality and Social Psychology Bulletin, 33*, 266–280.

Röhner, J. (2016). How to analyze (faked) implicit association test data by applying diffusion model analyses with the fast-dm software: A companion to Röhner & Ewers (2016). *The Quantitative Methods in Psychology, 12*, 220–231.

Röhner, J., & Ewers, T. (2016). Trying to separate the wheat from the chaff: Construct-and faking-related variance on the Implicit Association Test (IAT). *Behavior Research Methods, 48*, 243–258.

Röhner, J., Holden, R. R., & Schütz, A. (2022). Iat faking indices revisited: Aspects of replicability and differential validity. Behavior Research Methods, (pp. 1–24).

Röhner, J., & Lai, C. K. (2021). A diffusion model approach for understanding the impact of 17 interventions on the race implicit association test. *Personality and Social Psychology Bulletin, 47*, 1374–1389.

Röhner, J., Schröder-Abé, M., & Schütz, A. (2011). Exaggeration is harder than understatement, but practice makes perfect! Experimental Psychology, .

Röhner, J., Schröder-Abé, M., & Schütz, A. (2013). What do fakers actually do to fake the iat? an investigation of faking strategies under different faking conditions. *Journal of Research in Personality, 47*, 330–338.

Röhner, J., & Thoss, P. (2018). Ez: An easy way to conduct a more fine-grained analysis of faked and nonfaked implicit association test (iat) data. *The Quantitative Methods for Psychology, 14*, 17–35.

Röhner, J., & Thoss, P. J. (2019). A tutorial on how to compute traditional iat effects with r. *The Quantitative Methods for Psychology, 15*, 134–147.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin, 86*, 638.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic bulletin & review, 26*, 452–467.

Roy, V. (2020). Convergence diagnostics for markov chain monte carlo. *Annual Review of Statistics and Its Application, 7*, 387–412.

Rubin, M. (2020). Does preregistration improve the credibility of research findings? arXiv preprint http://arxiv.org/abs/2010.10513arXiv:2010.10513, .

Rubin, M., & Donkin, C. (2022). Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests. Philosophical Psychology, (pp. 1–29).

Ruiz, J. G., Andrade, A. D., Anam, R., Taldone, S., Karanam, C., Hogue, C., & Mintzer, M. J. (2015). Group-based differences in anti-aging bias among medical students. *Gerontology & Geriatrics Education, 36*, 58–78.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human perception and performance, 3*, 1.

Schimmack, U. (2021). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science, 16*, 396–414.

Schlenker, B. (1980). *Impression management: the self-concept social identity, and interpersonal relations*. Monterey: Brooks/Cole.

Schubert, A.-L., Hagemann, D., & Frischkorn, G. T. (2017). Is general intelligence little more than the speed of higher-order processing? *Journal of Experimental Psychology: General, 146*, 1498.

Schubert, A.-L., Hagemann, D., Voss, A., Schankin, A., & Bergmann, K. (2015). Decomposing the relationship between mental speed and mental abilities. *Intelligence, 51*, 28–46.

Schubert, A.-L., Nunez, M. D., Hagemann, D., & Vandekerckhove, J. (2019). Individual differences in cortical processing speed predict cognitive abilities: A model-based cognitive neuroscience account. *Computational Brain & Behavior, 2*, 64–84.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6*, 461–464.

Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and psychological measurement, 55*, 818–831.

Shieh, G. (2008). Improved shrinkage estimation of squared multiple correlation coefficient and squared cross-validity coefficient. *Organizational Research Methods, 11*, 387–407.

Shiffrin, R., Lee, M., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. *Cognitive Science: A Multidisciplinary Journal, 32*, 1248–1284.

Smith, C. T., & Nosek, B. A. (2011). Affective focus increases the concordance between implicit and explicit attitudes. Social Psychology, .

Smith, P. L. (2016). Diffusion theory of decision making in continuous report. *Psychological Review, 123*, 425–451.

Sokratous, K., Fitch, A., & Kvam, P. D. (2022). How to ask twenty questions and win: An automated model of risk preferences from small samples of willingness-to-pay prices. PsyArXiv, .

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76*, 485–493.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology), 64*, 583–639.

Stathakis, D. (2009). How many hidden layers and nodes? *International Journal of Remote Sensing, 30*, 2133–2147.

Stefanutti, L., Robusto, E., Vianello, M., & Anselmi, P. (2013). A discrimination-association model for decomposing component processes of the implicit association test. *Behavior research methods, 45*, 393–404.

Steffens, M. C. (2004). Is the implicit association test immune to faking? *Experimental psychology, 51*, 165–179.

Steingroever, H., Wabersich, D., & Wagenmakers, E.-J. (2020). Modeling across-trial variability in the wald drift rate parameter. Behavior Research Methods, (pp. 1–17).

Steyvers, M. (2011). MATJAGS 1.3: A matlab interface for JAGS .

Sun, R. (2008). Introduction to computational cognitive modeling. Cambridge handbook of computational psychology, (pp. 3–19).

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences, 24*, 94–95.

Thomas, D. R., & Zumbo, B. D. (2012). Difference scores from the point of view of reliability and repeated-measures anova: In defense of difference scores for data analysis. *Educational and Psychological Measurement, 72*, 37–43.

Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic bulletin & review, 21*, 227–250.

Turner, B. M., Van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: the neural drift diffusion model. *Psychological review, 122*, 312.

Turner, B. M., Wang, T., & Merkle, E. C. (2017). Factor analysis linking functions for simultaneously modeling neural and behavioral data. *NeuroImage, 153*, 28–48.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research, 37*, 141–188.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*, 550–592.

Vadillo, M. A., Malejka, S., Lee, D. Y., Dienes, Z., & Shanks, D. R. (2021). Raising awareness about measurement error in research on unconscious mental processes. Psychonomic Bulletin & Review, (pp. 1–23).

Van Dessel, P., Cummins, J., Hughes, S., Kasran, S., Cathelyn, F., & Moran, T. (2020). Reflecting on 25 years of research using implicit measures: Recommendations for their future use. *Social Cognition, 38*, s223–s242.

Verplanken, B., & Orbell, S. (2003). Reflections on past behavior: a self-report index of habit strength 1. *Journal of applied social psychology, 33*, 1313–1330.

Wagenmakers, E.-J., Van Der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy

Walczak, S., & Cerpa, N. (1999). Heuristic principles for the design of artificial neural networks. *Information and software technology, 41*, 107–117.

Wang, Z., Li, Y., Jin, Z., & Tamutana, T. T. (2019). How success enhances self-serving bias: A multinomial process model of the implicit association test. *Social Behavior and Personality: an international journal, 47*, 1–9.

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual review of psychology, 60*, 53–85.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*, 67–85.

Wrzus, C., Egloff, B., & Riediger, M. (2017). Using implicit association tests in age-heterogeneous samples: The importance of cognitive abilities and quad model processes. *Psychology and Aging, 32*, 432.

Xu, K., Nosek, B., & Greenwald, A. (2014). Psychology data from the race implicit association test on the project implicit demo website. Journal of Open Psychology Data, 2.

Yin, P., & Fan, X. (2001). Estimating r^2 shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education, 69*, 203–224.

Zhou, D.-X. (2020). Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis, 48*, 787–794.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision (pp. 19–27).