Behavior Research Methods (2020) 52:1640–1656
https://doi.org/10.3758/s13428-019-01340-z

Does discrimination beat association in the IAT? The discrimination-association model reconceived

Luca Stefanutti[1] · Egidio Robusto[1] · Michelangelo Vianello[1] · Pasquale Anselmi[1] · Anna Dalla Rosa[1] · Yoav Bar-Anan[2]

Published online: 11 March 2020

Abstract
The discrimination-association model (DAM; Stefanutti et al. 2013) disentangles two components underlying the responses to the implicit association test (IAT), which pertain to stimuli discrimination (the strength of the association of the stimuli with their own category) and automatic association (the strength of the association between targets and attributes). The assumption of the DAM that these two components sum into a single process generates critical drawbacks. The present work provides a new formulation of the model, called DAM-4C, in which stimuli discrimination and automatic association are separate, independent, and competing processes. Results of theoretical and simulation studies suggest that the DAM-4C outperforms the DAM. The IAT effect is found to vary with the association rates of the DAM-4C and not with those of the DAM. The parameters of the DAM-4C fitted on data from a Coca-Pepsi IAT are found to account for variance in brand attractiveness, taste preference, and cola choice that is not accounted for by the *D* score and the diffusion model. In addition, the association rates estimated on data from a Black-White IAT are in line with expectations.

Keywords Discrimination-association model · Poisson race model · Implicit Association Test · implicit measure · response time

This article presents a theoretical development of the discrimination-association model (DAM; Stefanutti et al., 2013) that is a formal model for decomposing the processes underlying the responses to the implicit association test (IAT; Greenwald et al., 1998), currently the most used and methodologically sound indirect measure of attitudes, stereotypes, and self-concept. At the heart of the IAT, there are two pairs of opposite categories, one consisting of target categories (e.g., *flowers* and *insects*) and the other consisting of attribute categories (e.g., *good* and *bad*), and a collection of stimuli which are exemplars of each of the categories (e.g., the picture of a tulip for the category flowers and the word "love" for the category *good*). The categories are displayed at the top-left and top-right screen corners, whereas the stimuli appear, one at a time, in the center of the screen. Participants have to categorize the stimuli into one of the categories by pressing, as quickly and accurately as possible, one of two response keys that are on the left and right sides of the keyboard. The procedure consists of seven blocks. Three are *practice* blocks, and involve the categorization of stimuli that represent either the target or the attribute categories. The remaining four are *test* blocks, and involve the categorization of stimuli representing the four categories. There are two response mappings that differ for the categories that are displayed on the same screen corner and, therefore, share the same response key. In the IAT under consideration, there is a mapping in which *flowers* shares the response key with *good*, and *insects* with bad, and a mapping in which *flowers* shares the response key with *bad*, and *insects* with *good*. When category pairs that are most strongly associated share the same response key (e.g., *flowers-good*; *insects-bad*), the mapping is called *compatible*. Otherwise, the mapping is called *incompatible*. Compatible and incompatible mappings for a certain individual are usually inferred by looking at latencies and accuracies of the responses (the compatible mapping is expected to be the one leading to faster and more accurate responses; Greenwald et al., 1998).

To date, the debate is open about the role and nature of the specific psychological processes behind the responses to the IAT. The closer we describe these psychological processes, the more precisely we can interpret and predict the outcome of the measure (Luce, 1996). A promising route in this direction is formal modeling. In this line of research, Klauer et al. (2007)

✉ Luca Stefanutti
luca.stefanutti@unipd.it

[1] Department FISPPA, University of Padua, Via Venezia 14, 35131 Padua, Italy

[2] School of Psychological Sciences, Tel-Aviv University , Tel-Aviv, Israel

proposed a diffusion model (DM) analysis of the IAT, whereas other authors proposed models specifically designed for the IAT, namely the Quad model (Conrey et al., 2005), the ReAL model (Meissner & Rothermund, 2013), and the aforementioned DAM (Stefanutti et al., 2013).

Based on the Poisson race model (PRM), the DAM deals with both response time and accuracy, and disentangles two process components pertaining to stimuli discrimination and automatic association. These components have interesting interpretations and provide substantive insights about the response process. Stimuli discrimination is interpreted as the strength of the association of the stimuli with their own category. Automatic association is interpreted as the strength of the association between targets and attributes. It is worth noting that also the Quad model and the ReAL model disentangle stimuli discrimination (parameter D in the Quad model, parameter L in the ReAL model) and automatic association (parameter AC in the Quad model, parameter A in the ReAL model). However, the two models analyze these components at the level of the entire IAT, whereas the DAM provides a more fine-grained analyses at the level of the different categories involved in the IAT. Being multinomial processing tree models, the Quad model and the ReAL model only account for response accuracy. Thus, they only use a very small part of the information provided by the IAT. In the same way as the DAM, the DM also accounts for both response time and accuracy. Parameters of the DM are a, z, v, t_0, η, s_z, and s_t. Parameter a is the respondent's speed–accuracy trade-off setting, with large values indicating slow and accurate responses. Parameter z measures response bias toward one of the two responses. Parameter v is the mean drift rate and quantifies the direction (toward correct or incorrect response) and the speed with which relevant information accumulates. Large values of v indicate both fast and accurate responses. Parameter t_0 reflects the nondecision component (e.g., encoding stimuli, motor response) of the response process. Parameters η, s_z, and s_t respectively quantify the variability of v, z, and t_0 across trials. A look at the parameters of the DM shows that this model does not provide any means of disentangling stimuli discrimination and automatic association.

In the DAM, the assumption was made that the stimuli discrimination and automatic association are merged into a single response process. We show that this assumption generates some drawbacks and provide a new formulation of the DAM in which stimuli discrimination and automatic association are separate and independent processes. Results of theoretical and empirical studies suggest that the new model outperforms the previous one in shedding light on the responses to the IAT.

The paper is organized as follows. After a brief overview of the PRM in Section "The Poisson race model", the DAM as developed by Stefanutti et al. (2013) is described in Section "The discrimination-association model". The new formulation of the DAM is presented in Section "The discrimination-association model reconceived". A theoretical comparison between the two models is the topic of Section "Theoretical comparison between the two models". The goodness of recovery of the parameters of the new model is investigated in Section "A goodness-of-recovery study". The predictive capabilities of the new formulation of the DAM and of the DM are compared in an empirical study presented in Section "Study 1: Comparison between DAM-4C and DM". An empirical validation of the parameters of the new formulation of the DAM is presented in Section "Study 2: Validation of DAM-4C parameters". Practical implications of the new model and suggestions for future research are explored and discussed in Section "Discussion".

The Poisson race model

The PRM is a stochastic process model belonging to the family of the so-called *counting models*, discussed in, for instance, Townsend and Ashby (1983). Such models assume that stimulus classification is an accumulation of "evidence" over time for each of the stimulus alternatives. As soon as the evidence for one of the alternatives exceeds a certain criterion, that alternative is chosen as the response.

A counting model consists of two or more random processes $N_i = \{N_i(t) : t \geq 0\}$, for $i \in Q$, named *counters*, each of which accumulates evidence for one of the categories in the set Q. This set usually contains two categories (e.g., *flowers* and *insects*), but in certain cases it can contain more. The whole classification process is a race among counters, so that the stimulus is classified as that alternative associated with the counter that accrues the required evidence in the shortest time.

A first standard assumption in counting models is that each counter N_i has its own termination criterion $K_i > 0$. In the discrete case (which is the one we assume here), criteria are integer parameters: the count of N_i must exceed a prespecified whole number K_i. A second standard assumption is that the counters operate independently and in parallel. In particular, independence means that increasing one counter does not affect the other(s). This assumption distinguishes such types of models from the so-called random-walk models, in which increasing one counter leads to a simultaneous decrease in each of the other counters.

The PRM arises as a consequence of the following specific assumptions:

(1) the inter-arrival times (i.e., times between successive counts in a counter N_i) are independent and identically distributed;
(2) the distribution of the inter-arrival times in N_i is exponential with rate λ_{ij} depending on the counter N_i and on

the presented simulus j. Thus the inter-arrival time density is

$$f(t) = \lambda_{ij} e^{-\lambda_{ij} t}.$$

By these two additional assumptions, the classification process turns out to be a race between two (or more) parallel and independent Poisson processes, and the overall random process $M = \{M(t) : t \geq 0\}$ is the minimum of the individual Poisson processes:

$$M = \min \{N_i : i \in Q\}.$$

Townsend and Ashby (1983) show that, in the two-category classification task $Q = \{a, b\}$, the joint density that, upon presentation of a stimulus of category a, the response is correct, and the time t is

$$f(t, a; a) = \frac{(\lambda_{aa} t)^{K_a} \lambda_{aa} e^{-\lambda_{aa} t}}{(K_a - 1)!} \sum_{k=0}^{K_b - 1} \frac{(\lambda_{ba} t)^k e^{-\lambda_{ba} t}}{k!},$$

whereas the joint density that the response is incorrect and the time is t is

$$f(t, b; a) = \frac{(\lambda_{ba} t)^{K_b} \lambda_{ba} e^{-\lambda_{ba} t}}{(K_b - 1)!} \sum_{k=0}^{K_a - 1} \frac{(\lambda_{aa} t)^k e^{-\lambda_{aa} t}}{k!}.$$

Interestingly, Poisson processes are a special type of continuous-time Markov chains with discrete state space (Ephraim & Mark, 2012). The states of the chain N_i are the nonnegative integer numbers less or equal to K_i, so that the state space of the chain is $\mathcal{S}_i = \{0, 1, \ldots, K_i\}$. Upon presentation of stimulus $j \in Q$, counter N_i behaves as a continuous-time Markov chain whose infinitesimal generator is the $(K_i + 1) \times (K_i + 1)$ square matrix

$$G_{ij} = \begin{pmatrix} -\lambda_{ij} & \lambda_{ij} & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_{ij} & \lambda_{ij} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_{ij} & \lambda_{ij} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

The chain starts in state 0 (no information accrued) and terminates in state K_i (all the required information has been accrued).

Counters N_i, as Markov processes, have the special property that the states from 0 to $K_i - 1$ of the chain are not observable (they are latent). Only the final state K_i is observable, as it may coincide with the emission of a response. Markov chains of this type are also known as *phase-type distributions* (Aldous & Shepp, 1987). It can be shown that, upon presentation of stimulus i, the minimum $\min\{N_i, N_j\}$ of the two phase-type distributed random variables N_i and N_j is still a phase-type distributed random variable (Buchholz et al., 2014). Defining $\mathcal{S}_i^* = \mathcal{S}_i \backslash \{K_i\}$, the state space of the minimum turns out to be the set

$$\mathcal{S}_{\min} = \left(\mathcal{S}_i^* \times \mathcal{S}_j^*\right) \cup \{i, j\}$$

that is the Cartesian product of the two state spaces, each reduced by one, plus the two absorbing states i and j.

As an example, the directed graph of the state space of a PRM with two counters N_a and N_b having identical termination criteria $K_a = K_b = 4$ is represented in Fig. 1. Each state is a pair (n_a, n_b) where n_a represents the number of counts of counter N_a, whereas n_b is the number of counts of counter N_b. The process starts in state (0, 0) and evolves, one count at the time in either of the two counters, until one of them exceeds four counts. This happens when one of the two absorbing states a or b is entered.

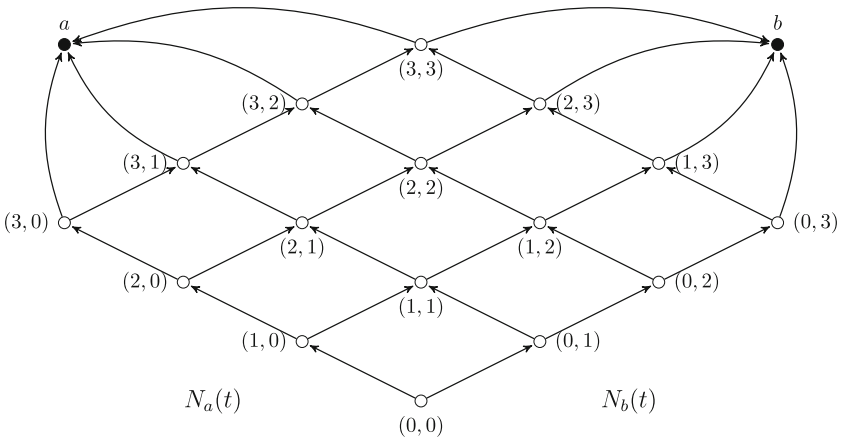

Fig. 1 State space of the Poisson race model with two counters, when it is seen as a finite state, continuous-time Markov chain. In this example, the two counters N_a and N_b have equal termination criteria $K_a = K_b = 4$. Each state of the Markov process is a pair (n_a, n_b) where n_a is the number of counts in counter N_a and n_b is the number of counts in N_b. White circles represent transient states, whereas black circles represent absorbing states.

The discrimination-association model

This section describes the DAM as developed by Stefanutti et al. (2013). Applications of the model to empirical data have been described in Anselmi et al. (2013) and in Stefanutti et al. (2013). An user-friendly application for fitting the model to IAT data has been developed by Stefanutti et al. (2014).

Let $Q = \{a, b, +, -\}$ be the set of the four categories of the IAT, with a and b being the target categories, and + and − being the attribute categories. The stimuli are exemplars of each of the four categories in Q.

The DAM assumes the existence of four counters N_i, one for each of the four categories in Q. Once a stimulus is presented on the screen, each counter starts accumulating selective evidence about a specific characteristic of it. For example, counter N_a accumulates evidence about the membership of the stimulus to category a. The four counters are assumed to behave as Poisson processes.

Model parameters are the rates at which evidence is accumulated on each counter, and the termination criteria. There is a rate for each pair that can be formed by taking one of the four counters N_a, N_b, N_+, N_- and one of the four categories a, b, +, −. According to this formulation, for $i, j \in \{a, b, +, -\}$, the parameter λ_{ij} is the average amount of evidence that counter N_i accumulates, in the time unit, when a stimulus of category j is presented on the screen. For example, λ_{aa} and λ_{+a} represent the amount of evidence respectively accumulated by counters N_a and N_+ when the presented stimulus belongs to category a.

Table 1 displays the 16 rates of the DAM. They can be grouped into discrimination rates and association rates. The discrimination rates regard the amount of evidence that target (respectively, attribute) categories accumulate when target (respectively, attribute) stimuli are presented. In particular, the four rates λ_{aa}, λ_{ab}, λ_{ba}, λ_{bb} (upper left 2 × 2 submatrix of the table) are involved in the discrimination between a and b, whereas the four rates λ_{++}, λ_{+-}, λ_{-+}, λ_{--} (lower right 2 × 2 submatrix) are involved in the discrimination between + and −. A fundamental requirement of the IAT procedure is that the stimuli are prototypical exemplars of their own category (Lane et al., 2007; Nosek et al., 2007a). The discrimination rates provide information about satisfaction of this requirement by the selected stimuli. For instance, the rates λ_{aa} and λ_{ba} are respectively involved in the correct and incorrect discrimination of stimuli a. The former rate being much larger than the latter suggests that the stimuli chosen to represent the category a suit this purpose well.

Table 1 Discrimination rates (upper left and lower right **2 × 2** matrices) and association rates (lower left and upper right **2 × 2** matrices)

	Stimulus categories			
Counters	a	b	+	−
N_a	λ_{aa}	λ_{ab}	λ_{a+}	λ_{a-}
N_b	λ_{ba}	λ_{bb}	λ_{b+}	λ_{b-}
N_+	λ_{+a}	λ_{+b}	λ_{++}	λ_{+-}
N_-	λ_{-a}	λ_{-b}	λ_{-+}	λ_{--}

The association rates concern the amount of evidence that target (respectively, attribute) categories accumulate when attribute (respectively, target) stimuli are presented. The four rates λ_{+a}, λ_{+b}, λ_{-a}, λ_{-b} (lower left 2 × 2 submatrix) are the rates at which evidence concerning membership to attribute categories are accumulated when a target stimulus is presented (target-driven associations). The four rates λ_{a+}, λ_{a-}, λ_{b+}, λ_{b-} (upper right 2 × 2 submatrix) are the rates at which evidence concerning membership to target categories are accumulated when an attribute stimulus is presented (attribute-driven associations). The association rates are very informative in practical applications of the model: The particular values taken by these parameters might enable the identification of patterns of automatic association between targets and attributes that differ from one individual to another in both nature and meaning. The D score (Greenwald et al., 2003) is an effect size measure that quantifies the difference between the performance of the respondent in the two types of test blocks. A positive D score in a Flowers-Insects IAT, for instance, might indicate an implicit preference for flowers over insects. However, no information is provided on the meaning of this preference. The investigation of the values of the association rates would serve this purpose. For instance, they could reveal that the individual (*a*) likes flowers and is indifferent to insects, *(b)* dislikes insects and is indifferent to flowers, (*c*) likes flowers and dislikes insects, (*d*) likes flowers more than insects, or (*e*) dislikes insects more than flowers.

The termination criteria vary across block types. Thus, there are three termination criteria, one for the practice blocks K_P, one for the compatible blocks K_C, and one for the incompatible blocks K_I (P, C, and I stand for *practice*, *compatible*, and *incompatible*, respectively). These parameters can be interpreted as either task difficulty or individual cautiousness. Since the test blocks (compatible and incompatible) are a double classification task, they are expected to be more difficult than the practice blocks. Moreover, when all termination criteria are taken into account, the following inequalities are expected:

$$K_P < K_C < K_I.$$

We now describe which counters and model parameters are involved in the emission of the observable responses in the different block types of the IAT. In the practice blocks involving targets, a race takes place between the two Poisson processes $N_a = \{N_a(t) : t \geq 0\}$ and $N_b = \{N_b(t) : t \geq 0\}$, with termination criterion K_P. Supposing that a stimulus of category a is presented on the screen, the rate of N_a is λ_{aa} and the rate of

N_b is λ_{ba}. The random process is the minimum between N_a and N_b (i.e., $\min\{N_a, N_b\}$). In the practice blocks involving attributes, the race is between $N_+ = \{N_+(t) : t \geq 0\}$ and $N_- = \{N_-(t) : t \geq 0\}$, and the random process is $\min\{N_+, N_-\}$.

In the test blocks, target and attribute categories share the response key. Suppose that the compatible blocks map *a* and + to the left key, and *b* and − to the right key. Concerning the relationship between each of the four processes and each of the two observable responses, it can be seen that processes N_a and N_+ both produce the response *left*, whereas processes N_b and N_- both produce the response *right*. A basic assumption of the DAM is that evidence units that contribute to the same observable response are accumulated by the same compound process. According to a well-known property of Poisson processes, merging two independent Poisson processes results in another Poisson process with rate equal to the sum of the individual rates. In the case under consideration, the race is between the two compound processes $N_a + N_+$ and $N_b + N_-$, with termination criterion K_C. Assuming that a stimulus of category *a* is presented on the screen, the rate of $N_a + N_+$ is $\lambda_{aa} + \lambda_{+a}$ whereas the rate of $N_b + N_-$ is $\lambda_{ba} + \lambda_{+a}$. Thus, the overall random process is $\min\{(N_a + N_+), (N_b + N_-)\}$.

Assume now that the incompatible blocks map *b* and + to one key, and *a* and − to the other key. In these blocks, the race is between the two compound processes $N_b + N_+$ and $N_a + N_-$, with termination criterion K_I. Supposing that a stimulus of category *a* is presented, the rate of $N_b + N_+$ is $\lambda_{ba} + \lambda_{+a}$, whereas the rate of $N_a + N_-$ is $\lambda_{aa} + \lambda_{-a}$. Then, the overall random process is $\min\{(N_b + N_+), (N_a + N_-)\}$. Thus, regardless of the particular type of blocks, the model under consideration assumes that the race is always between two Poisson processes.

The discrimination-association model reconceived

Suppose that a target stimulus of category *a* has just appeared on the screen in one of the two test blocks of the IAT. A decision mechanism is modeled, consisting of two separate processes: a discrimination process $D = \{D(t) : t \geq 0\}$ and an association process $A = \{A(t) : t \geq 0\}$. Upon stimulus presentation, the two processes operate in parallel, and a simplifying assumption is that they are also independent.

The task of *D* is to discriminate the stimulus presented on the screen as belonging to either category *a* or category *b*. This is a two-choice decision task and there are, essentially, two classes of models of accuracy and response times for this type of tasks: the counting models (e.g., the PRM) and the random walk models (e.g., the diffusion model; see, e.g., Townsend & Ashby, 1983). In this work, the process *D* is modeled as a race between two Poisson processes $D_a = \{D_a(t) : t \geq 0\}$ and $D_b = \{D_b(t) : t \geq 0\}$ with rates varying with both counter type (*a* or *b*) and stimulus category (*a* or *b*). A single termination criterion K_D, equal across counters and stimuli, is assumed. It should be observed that whenever higher complexity is needed, it is always possible to drop the assumption of a single criterion. Valid alternatives could be that the termination criterion varies with the counters, or with the stimuli or both. Such alternatives are not considered here. Then we have $D = \min\{D_a, D_b\}$. The parameters of this process are the four discrimination rates λ_{aa}, λ_{ab}, λ_{ba}, λ_{bb}, and the termination criterion K_D (it should be observed that in the notation λ_{ij}, *i* is the counter and *j* is the stimulus). The two parameters λ_{aa}, λ_{bb} are named the *correct discrimination rates*, whereas λ_{ab}, $\lambda_{ba}$are named the *incorrect discrimination rates*. The parameter K_D is named the *criterion of the discrimination process*.

The association process *A* associates the stimulus presented on the screen to one of two opposed evaluative attributes (usually *good* and *bad*). This is also a two-choice decision task, though the "decision" is regarded here as the outcome of an automatic process. The process *A* is also modeled as a race between two Poisson processes with rates varying with both stimulus category (*a* or *b*) and counter type ("+" for *good* and "–" for *bad*). A single termination criterion K_A, equal across counters and stimuli, is assumed (but the considerations here are similar to those already done for the discrimination process). Let $A_+ = \{A_+(t) : t \geq 0\}$ be the Poisson process representing the counter for category +, and $A_- = \{A_-(t) : t \geq 0\}$ be the Poisson process representing the counter for category −. Then, *A* is the minimum between A_+ and A_-, that is $A = \min\{A_+, A_-\}$. The parameters of this process are the four association rates λ_{+a}, λ_{+b}, λ_{-a}, λ_{-b}, and the termination criterion K_A.

The overall random process $M = \{M(t) : t \geq 0\}$ is a race between the discrimination process *D* and the association process *A*, meaning that $M = \min\{D, A\}$. It follows at once that $M = \min\{D_a, D_b, A_+, A_-\}$ is a race among four Poisson processes, two of which are discrimination processes, whereas the other two are association processes. Therefore, the overall model turns out to be a Poisson race model with four parallel counters. This is the *discrimination-association model with four counters* (DAM-4C). Upon presentation of a stimulus on the screen, the counter that first exceeds its criterion generates the response (either *left* or *right*) which is associated with it. In particular, once a counter has terminated, the corresponding response is selected according to a deterministic mapping that depends on how the four categories *a*, *b*, + and − are mapped to the left and right keys.

If a stimulus of category $i \in \{a, b\}$ appears on the screen, the probability density that the process D_i finishes first at time $t > 0$ is

$$g_i(t) = f_i(t)F_j(t)F_+(t)F_-(t)$$

where

$$f_i(t) = \frac{\lambda_i(\lambda_i t)^{K_D-1} e^{-\lambda_i t}}{(K_D-1)!}$$

is the density of counter D_i exceeding its criterion K_D at time t, whereas

$$F_j(t) = \sum_{k=0}^{K_D-1} \frac{(\lambda_j t)^k e^{-\lambda_j t}}{k!},$$

$$F_+(t) = \sum_{k=0}^{K_A-1} \frac{(\lambda_+ t)^k e^{-\lambda_+ t}}{k!},$$

$$F_-(t) = \sum_{k=0}^{K_A-1} \frac{(\lambda_- t)^k e^{-\lambda_- t}}{k!}.$$

Similarly, the probability density for process A_+ finishing first at time $t > 0$ is

$$g_+(t) = f_+(t)\overline{F}_-(t)\overline{F}_i(t)\overline{F}_j(t).$$

Suppose that, in the ongoing experimental block, both *i* and + are mapped on the same key, and the stimulus *i* is presented on the screen. Then, the probability density of a correct (1) response at time t is

$$\begin{aligned} g_1(t) &= g_i(t) + g_+(t) \\ &= \left(f_i(t)F_+(t) + f_+(t)F_i(t)\right)F_j(t)F_-(t), \end{aligned}$$

whereas the density of an incorrect (0) answer at time t is

$$\begin{aligned} g_0(t) &= g_j(t) + g_-(t) \\ &= \left(f_j(t)F_-(t) + f_-(t)F_j(t)\right)F_i(t)F_+(t). \end{aligned}$$

As already observed with the Poisson race model, even in the DAM-4C, the overall process M can be regarded as a continuous-time Markov chain whose state space is

$$\mathcal{S}_M = \left(\mathcal{S}_i^* \times \mathcal{S}_j^* \times \mathcal{S}_+^* \times \mathcal{S}_-^*\right) \cup \{a, b, +, -\}.$$

Application of the DAM-4C to the IAT

In the application of the DAM-4C to the IAT, additional constraints need be considered. The IAT is roughly divided into three types of blocks: the practice blocks, the compatible blocks, and the incompatible blocks. Since the practice blocks are a single classification task, it seems reasonable to assume that the association process A is quiescent all along these blocks, and that only the discrimination process D is active. Hence, in the practice blocks, the DAM-4C behaves as a standard two-counter PRM with a single termination criterion K_{DP} (*DP* stands for *discrimination in practice* blocks), assumed to be equal across stimulus categories and counters.

In the compatible and incompatible test blocks, both processes D and A are active. Since in these blocks the task is a double classification, its difficulty increases in comparison with the practice blocks. Because of the increased difficulty, it would be unrealistic to expect that the parameters governing the discrimination in the test blocks are the same as those in the practice blocks. It is assumed that the correct and incorrect discrimination rates are not affected by task difficulty. Thus, the only parameter of the process D that is affected by an increase in task difficulty is the termination criterion. Thus, the model has two separate discrimination criteria: a criterion K_{DP} in the practice blocks, and a (possibly) different criterion K_{DT} in the test blocks (*DT* stands for *discrimination in test* blocks). In all other respects, the parameters of the process D are exactly the same across the compatible and incompatible blocks, so that any systematic differences in latency or accuracy between the two types of test blocks cannot be ascribed to the discrimination process.

The four rates of the association process A are also assumed to be invariant across test blocks. In principle, these four parameters should provide a measure of how fast the association process is, regardless of the test blocks in which it operates. Given these constraints, the only parameter that can vary across test blocks is the criterion of A. We assume two different criteria for this process: a criterion K_{AC} in the compatible blocks, and a different criterion K_{AI} in the incompatible blocks (*AC* and *AI* stand for *association in compatible* blocks and *association in incompatible* blocks, respectively).

We do not incorporate into the model any inequality constraints concerning the discrimination rates, association rates, or termination criteria. However, some inequalities can be postulated, which could then be verified empirically. In the first place, the double classification of the test blocks should increase the task difficulty relative to the single classification task of the practice blocks. If this is true, one should expect $K_{DP} < K_{DT}$.

Furthermore, whenever it is clear which one of the two types of test blocks is the compatible one (e.g., in a Flowers-Insects IAT, the compatible blocks are expected to be those in which pictures of flowers and positive attributes are mapped to the same key), some hypotheses concerning the direction of the inequality between the two association criteria K_{AC} and K_{AI} can be formulated. Consider a Flowers-Insects IAT and suppose first that the two criteria are equal. In the incompatible blocks, flower images and positive attributes are mapped to different keys. If the process A_+ is fast enough with *flower* stimuli (due to a strong association), it can win the race rather often, leading to frequent inaccuracies. Hence, elevating the criterion in the incompatible task can be seen as a way of counteracting inaccuracy due to a high association rate. On the contrary, in the compatible task the two stimuli are mapped to the same key, and the fast responses of A_+ increase, rather than reduce, accuracy. Hence, in these blocks, a smaller

criterion works well. These considerations lead us to formulate the hypothesis that $K_{AC} < K_{AI}$.

Theoretical comparison between the two models

As a consequence of the different number of counters, the two models differ in structure, and make markedly different predictions of latency and accuracy in the IAT. The DAM can be regarded as the minimum of the sum of Poisson processes. Its form varies depending on the test blocks. Assuming that a target stimulus is presented on the screen, in one type of test block it takes on the form

$$\min\{(D_a + A_+), (D_b + A_-)\},$$

whereas in the other type of test block it becomes

$$\min\{(D_b + A_+), (D_a + A_-)\}.$$

Because of the sum, there is no real separation between the discrimination and the association processes in the DAM. Instead, they are collapsed into the same Poisson process, sum of the two. A consequence of this fact is that discrimination and association will always share the same termination criterion, irrespective of the blocks in which they operate.

Another rather important point concerning the DAM is that the two processes A_+ and D_i, as well as A_- and D_j, are not independent, indeed. More precisely, given a certain termination criterion K, the probability of having $k > 0$ counts in A_- is not independent of the number of counts in D_i. For instance, if at time $t > 0$, $D_i(t) < K - 1$, then there is a positive probability that $A_+(t) = 2$. However, if $D_i(t) = K - 1$, then the probability of $A_+(t) = 2$ must be zero, for otherwise one would have $D_i(t) + A_+(t) > K$.

The form of the DAM-4C does not depend on the block, as it is simply the minimum of the four processes in both types of test blocks:

$$\min\{D_i, D_j, A_+, A_-\}.$$

There is independence among processes in this model. Moreover, A and D are really separate processes operating independently and in parallel. As such, they can have different termination criteria.

Numerical example

A numerical example can help in better appreciating the different behaviors of the two models under equivalent conditions. We examine how the mean latency and accuracy predicted by each of the two models vary as functions of certain model parameters. With $Q = \{a, b, +, -\}$, suppose that the compatible task maps a and + to the same key, and that a stimulus of class a has appeared on the screen. We first examine how the predicted mean latencies and response probabilities vary as the association rate λ_{+a} varies, keeping constant all the other parameters. In particular, for the DAM, let $\lambda_{aa} = 1.0$, $\lambda_{ba} = 0.4$, $\lambda_{-a} = 0.01$, $K_C = 5$, and $K_I = 8$. The latency curves for this choice of model's parameters are given in Fig. 2.

The figure consists of three panels. The upper left panel shows the latency curves for the correct answer, incorrect answer, and average in the compatible blocks; the upper right panel shows the same three curves, but for the incompatible blocks; finally, in the bottom panel, each curve is obtained as the difference of a mean latency curve in the incompatible task minus the corresponding mean latency curve in the compatible task. This difference is the "delay" that one observes when comparing reaction times in the incompatible blocks with those in the compatible blocks. It can be regarded as an IAT effect. The bottom panel is also the most instructive one. It shows the following rather odd behavior of the DAM: when λ_{+a} is zero, the delay is about 150 milliseconds. This delay increases slightly until λ_{+a} reaches the value of about 0.4, and then it starts decreasing monotonically: the higher the association rate, the smaller the IAT effect. This goes against the expectation that a positive relationship exists between the association rate and the IAT effect.

The reason for this behavior is that, as λ_{+a} increases, both the mean latency in the incompatible task (upper right panel of Fig. 2) and the accuracy in the same task (Fig. 3) decrease. What the model essentially predicts is that a higher association rate makes the respondent faster and less accurate in the incompatible blocks.

Figure 4 shows what happens with the DAM-4C. The parameters of the model were $\lambda_{aa} = 1.0$, $\lambda_{ba} = 0.4$, $\lambda_{-a} = .01$, $K_{DT} = K_{AC} = 5$, and $K_{AI} = 8$. The bottom panel of the figure shows the delay curves. Unlike the DAM, when the association rate is zero, the IAT effect is also zero. Then the curve monotonically increases up to a maximum for a λ_{+a} of about 1.8, after which it starts decreasing at a slower rate. In the DAM-4C, as the association rate increases, even the IAT effect increases. However, when the association rate becomes high, the IAT effect goes down again. The reason is given by the top right panel of Fig. 5. There we see that accuracy markedly decreases for values of λ_{+a} greater than 1. Again, as already seen with the DAM, the model says that a too high association rate makes the respondent faster and more inaccurate. The difference with the DAM is that, in this model, this happens only when the association rate is too high relative to the criterion K_{AI} used.

We now examine how the predicted mean latencies vary with the termination criterion K_I, keeping constant all the other parameters. For the DAM, let $\lambda_{aa} = .30$, $\lambda_{ba} = .15$, $\lambda_{+a} = .50$, $\lambda_{-a} = .01$, and $K_C = 5$, and let K_I vary from 5 to 25. The latency curves for this choice of the model parameters are displayed in

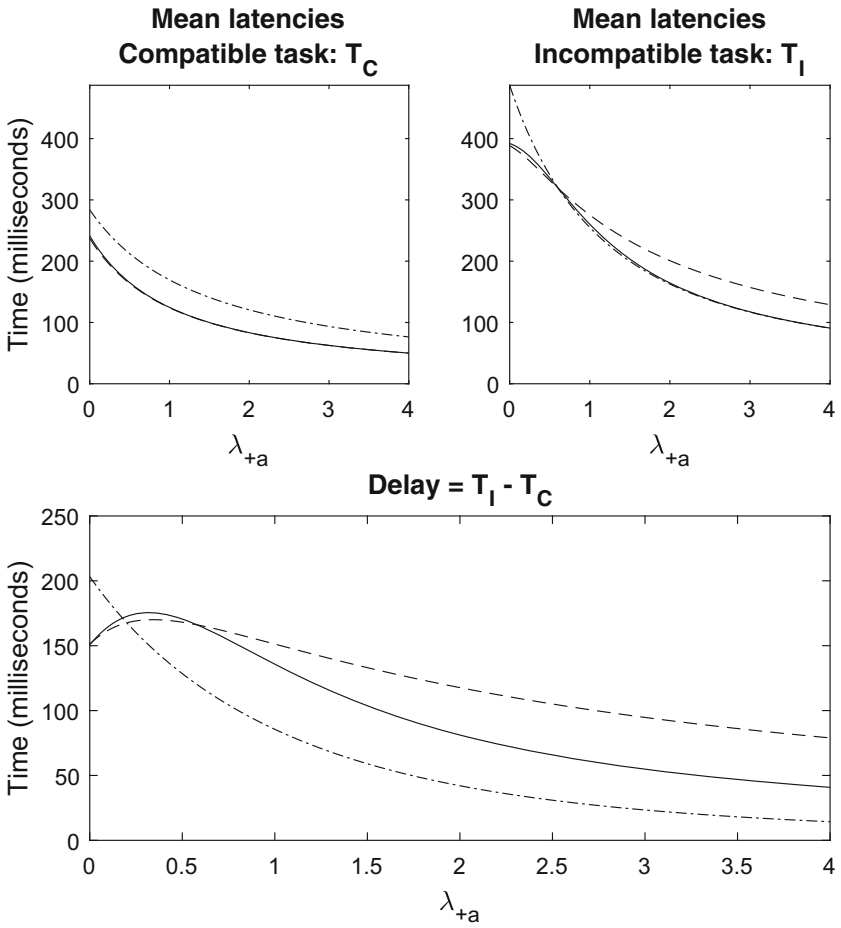

Fig. 2 Mean latency as a function of the λ_{+a} parameter in the DAM model. Dashed curves: correct answer; dash-dotted curves: incorrect answer; solid curves: average. Upper left panel: compatible task; upper right panel: incompatible task; lower panel: delay measured as difference between the mean latency in the incompatible task minus the mean latency in the compatible task.

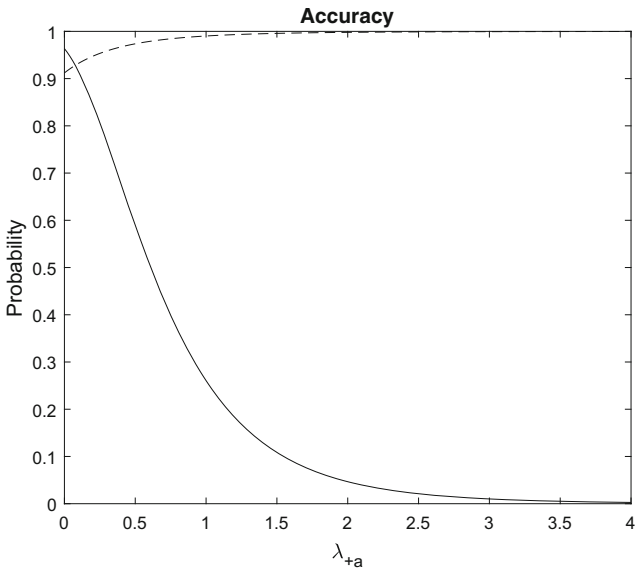

Fig. 3 Accuracy as a function of the λ_{+a} parameter in the DAM model. Dashed curve: compatible task; solid curve: incompatible task.

Fig. 6. As expected, the variation of K_I does not affect the mean latency in the compatible task (upper left panel of Fig. 6), whereas it influences the mean latency in the incompatible task (upper right panel of Fig. 6) that *linearly* increases with the value of K_I. Interestingly, there is a positive linear relationship between the delay that one observes when comparing reaction times in the incompatible blocks with those in the compatible blocks and the difference between the two termination criteria K_I and K_C. It should be observed that such a linear dependence is of a deterministic nature, and hence the two quantities at issue are perfectly correlated in the model. Stated differently, up to a linear transformation, the difference $K_I - K_C$ and the difference between expected reaction times in the two blocks are essentially the same thing for the DAM. Given this, if the model fits the data well, then it is likely to have small to negligible residuals between the observed mean reaction times and those expected by the model. In such cases, $K_I - K_C$ would account for the greatest part of the variance of the observed mean response time difference, leaving little to the remaining parameters of the model. With a perfect fit, $K_I - K_C$ would be the only predictor of the observed mean response time difference.

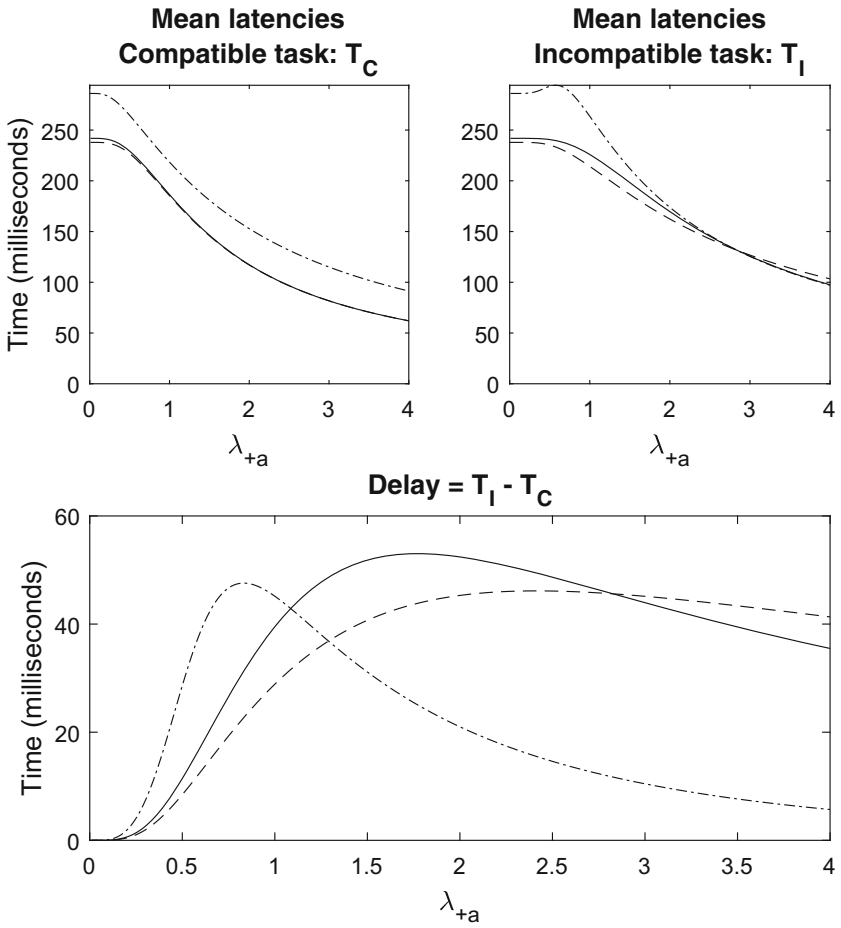

Fig. 4 Mean latency as a function of the λ_{+a} parameter in the DAM-4C model. Dashed curves: correct answer; dash-dotted curves: incorrect answer; solid curves: average. Upper left panel: compatible task; upper right panel: incompatible task; lower panel: delay measured as difference between the mean latency in the incompatible task minus the mean latency in the compatible task.

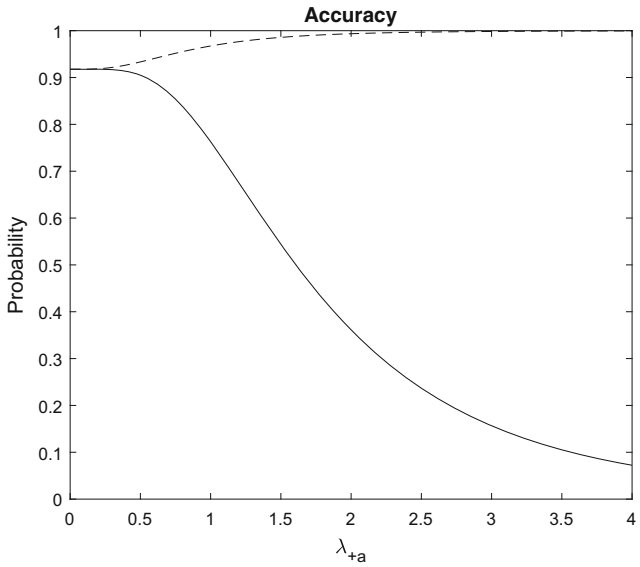

Fig. 5 Accuracy as a function of the λ_{+a} parameter in the DAM-4C model. Dashed curve: compatible task; solid curve: incompatible task.

This result, together with that related to the variation of the association rate λ_{+a}, explains the results found by Stefanutti et al. (2013) in an empirical application of the model to the data of a Coca-Pepsi IAT. A total of nine contrast measures were computed for each respondent to the IAT, four based on the estimates of the discrimination rates, four based on the estimates of the associations rates, and one based on the estimates of the two termination criteria K_C and K_I. The nine measures were the independent variables of a regression analysis in which the D score (Greenwald et al., 2003), which is a very common measure of the IAT effect, was the dependent variable. The difference $K_I - K_C$ was, by far, the strongest predictor of the D score, whereas the four contrast measures based on the association rates were significant yet much less strong.

Figure 7 shows what happens with the DAM-4C when the parameters are: $\lambda_{aa} = .30$, $\lambda_{ba} = .15$, $\lambda_{+a} = .50$, $\lambda_{-a} = .01$, and $K_{DT} = K_{AC} = 5$, and K_{AI} varies from 5 to 25. The delay is 0 milliseconds when $K_{AI} = K_{AC} = 5$. It increases until K_{AI} reaches about 20, and then it remains constant. Thus, unlike the DAM, in this model the relation between K_{AI} and the delay

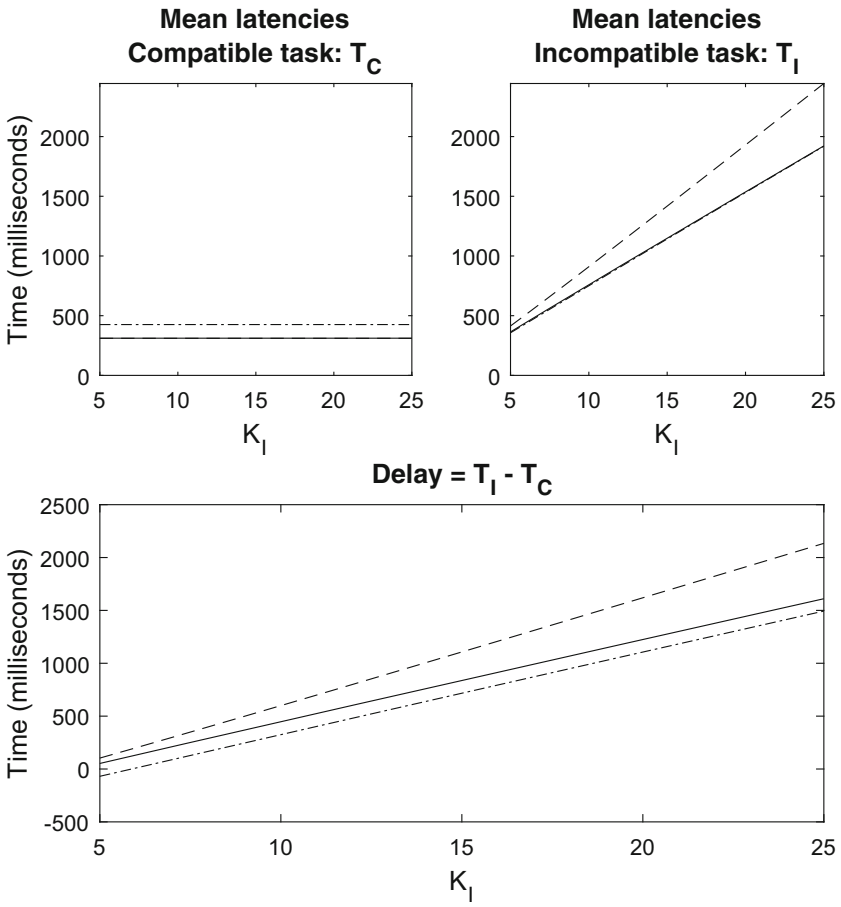

Fig. 6 Mean latency as a function of the K_I parameter in the DAM model. Dashed curves: correct answer; dash-dotted curves: incorrect answer; solid curves: average. Upper left panel: compatible task; upper right panel: incompatible task; lower panel: delay measured as difference between the mean latency in the incompatible task minus the mean latency in the compatible task.

in reaction times is not linear, and for large values of K_{AI}, the delay approaches a constant value.

A goodness-of-recovery study

This section aims to investigate the recovery of DAM-4C parameters in IATs of typical length (200 trials; Nosek et al., 2007a). Two conditions are simulated that differ for the contribution of the discrimination and association processes to the IAT effect. Suppose that the compatible task maps "*a*" and "+" to one response key, and "*b*" and "–" to the other response key. In Condition A, the parameters used for simulating latencies and accuracies of the responses were $\lambda_{aa} = \lambda_{bb} = \lambda_{++} = \lambda_{--} = 4.00$, $\lambda_{ba} = \lambda_{ab} = \lambda_{-+} = \lambda_{+-} = 0.50$, $\lambda_{+a} = \lambda_{a+} = \lambda_{-b} = \lambda_{b-} = 4.00$, $\lambda_{-a} = \lambda_{a-} = \lambda_{+b} = \lambda_{b+} = .05$, $K_{DP} = 20.00$, $K_{DT} = K_{AC} = 25.00$, and $K_{AI} = 30.00$. Thus, Condition A represents a situation in which:

(a) the λ parameters pertaining to the correct discrimination of each stimulus category are eight times larger than those pertaining to the incorrect discrimination;

(b) the λ parameters pertaining to the associations of "*a*" with "+" and "*b*" with "–" are eight times larger than those pertaining to the associations of "*a*" with "–" and "*b*" with "+";

(c) The probability of the responses in the compatible blocks is equally influenced by the discrimination and association processes (the correct discrimination rates are equal to the association rates involved in the compatible blocks, and the termination criterion of the discrimination process in the test blocks K_{DT} is equal to that of the association process in the compatible blocks K_{AC}).

Condition B differs from Condition A for only the parameters λ_{+a}, λ_{a+}, λ_{-b}, and λ_{b-} which are all set to be equal to 1.00. Thus, in Condition B, the observable response is more often determined by the discrimination process (which terminates earlier) than by the association process (which terminates later): The termination criterion of the discrimination process in the test blocks K_{DT} was equal to that of the association process in the compatible blocks K_{AC} (25.00), but the correct discrimination rates were larger (4.00) than the association rates involved in the compatible blocks (1.00).

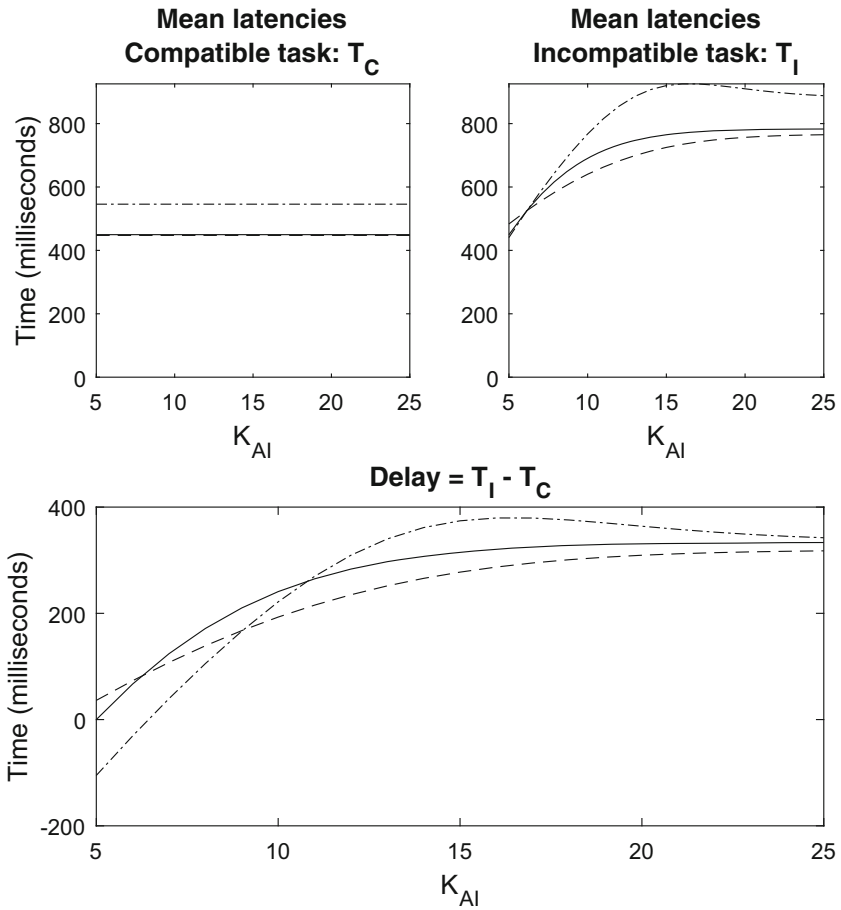

Fig. 7 Mean latency as a function of the K_{AI} parameter in the DAM-4C model. Dashed curves: correct answer; dash-dotted curves: incorrect answer; solid curves: average. Upper left panel: compatible task; upper right panel: incompatible task; lower panel: delay measured as difference between the mean latency in the incompatible task minus the mean latency in the compatible task.

The responses to 100 IATs were simulated for each of the two conditions, and the DAM-4C was estimated on each of them. Table 2 shows medians, means, and standard deviations of the parameter estimates. In Condition A, the estimates of all the parameters are close to the corresponding true values and rather consistent across the 100 simulated IATs. In Condition B, the estimates of the rates and the termination criteria involved in the association process are far from the true values and largely variant across the simulated IATs, whereas those of the other parameters are consistent and close to the true values. This result suggests that, when the association process contributes little to the responses to the IAT, the 200 trials of the typical IAT procedure may not be sufficient to provide reliable estimates of the parameters involved in this process.

Study 1: Comparison between DAM-4C and DM

The present study aims to compare the predictive capabilities of DAM-4C and DM. These two models were estimated based on the responses of 199 psychology students (M_{age} = 23.66, *SD* = 2.55; 122 females) at the University of Padua to a Coca-Pepsi IAT. This data set was already analyzed by Stefanutti et al. (2013) via the DAM.

Participants were tested individually in a laboratory. They were first presented with the Coca-Pepsi IAT according to the structure in Table 3. Ten color brand pictures were used to represent the target categories *Coca Cola* and *Pepsi Cola*, and 16 words were used to represent the attribute categories *good* (glory, good, happiness, joy, laughing, love, peace, pleasure) and *bad* (annoying, bad, evil, failure, hate, horrible, pain, terrible). The stimuli were presented in the center of the computer screen in an alternating fashion, and participants were asked to categorize them by pressing, as quickly and accurately as possible, the response key "Q" or "P". A red "X" appeared in case of a mistake, and it disappeared after the correct response was given.

Participants were then presented with two dichotomous questions asking them which was, in their opinion, the better-tasting cola and more attractive brand between Coca Cola and Pepsi Cola. At the end, participants were invited to choose between a free can of Coca Cola or Pepsi Cola, which was offered to them as a reward for their participation in the

Table 2 Goodness of recovery of DAM-4C parameters

	Condition A				Condition B			
Parameter	True	Median	Mean	*SD*	True	Median	Mean	*SD*
λ_{aa}	4.00	4.13	4.21	0.57	4.00	4.03	4.09	0.44
λ_{bb}	4.00	4.12	4.24	0.59	4.00	4.05	4.10	0.44
λ_{++}	4.00	4.20	4.25	0.57	4.00	4.02	4.08	0.48
λ_{--}	4.00	4.12	4.25	0.60	4.00	4.08	4.10	0.44
λ_{ba}	0.50	0.29	0.41	0.46	0.50	0.23	0.31	0.27
λ_{ab}	0.50	0.30	0.42	0.43	0.50	0.32	0.34	0.25
λ_{-+}	0.50	0.41	0.50	0.48	0.50	0.31	0.34	0.26
λ_{+-}	0.50	0.34	0.41	0.40	0.50	0.30	0.34	0.29
λ_{+a}	4.00	4.48	4.96	2.14	1.00	213.68	1.09×10^5	2.62×10^5
λ_{-b}	4.00	4.68	5.05	2.12	1.00	10.26	5.76×10^4	1.35×10^5
λ_{-a}	0.50	0.34	0.83	1.44	0.50	151.81	9.74×10^4	2.38×10^5
λ_{+b}	0.50	0.25	0.66	1.10	0.50	48.92	8.76×10^4	1.81×10^5
λ_{a+}	4.00	4.56	5.09	2.22	1.00	58.94	9×10^4	2.32×10^5
λ_{b-}	4.00	4.60	5.15	2.22	1.00	74.15	9.46×10^4	2.73×10^5
λ_{b+}	0.50	0.38	0.87	1.35	0.50	247.56	1.00×10^5	2.08×10^5
λ_{a-}	0.50	0.33	0.65	0.90	0.50	30.92	1.01×10^5	2.43×10^5
K_{DP}	20.00	20.77	21.07	2.76	20.00	20.39	20.48	2.21
K_{DT}	25.00	25.76	26.36	3.54	25.00	25.64	25.70	2.80
K_{AC}	25.00	28.49	35.45	20.08	25.00	1.31×10^5	2.33×10^6	3.90×10^6
K_{AI}	30.00	34.28	38.66	14.16	30.00	1.31×10^5	2.06×10^6	3.18×10^6

study. The two cans were disposed on a table, at the same distance from the participants. The experimenter registered the choice of the participants after they had left the laboratory.

The DAM-4C and the DM were estimated based on the data of each participant. Typically, the latency of the incorrect responses is increased by the time that is required to correct them (see, e.g., Greenwald et al., 2003). We used the latency of the incorrect responses, and discarded the time needed to correct them. The DAM-4C was estimated on the data from all seven IAT blocks using MATLAB functions written by the first author (available upon request). These functions compute maximum-likelihood estimates of the parameters of the DAM-4C. Since the maximum-likelihood estimation is sensitive to outliers and contaminants in the distributions of response times (Ratcliff & Tuerlinckx, 2002), Tukey's criterion (see, e.g., Hoaglin et al., 1983) was used for discarding the trials with outlying latencies. This led to the deletion of 6.84% of all the responses (from 1.63% to 14.67% per participant). Following Studies 2 and 3 (Klauer et al., 2007), two DMs were estimated for each participant, one on the data from the two blocks *Pepsi-bad/Coca-good* and the other on the data from the two blocks *Coca-bad/Pepsi-good*. Parameter z was set equal to $a/2$, that is, equal to the position corresponding to the absence of response bias (Klauer et al., 2007). Maximum-likelihood estimates of the parameters of the DM were computed using the software Fast-DM (Voss & Voss, 2007, 2008).

Table 3 Structure of the Coca-Pepsi IAT

Block type	No. of trials	Left labels	Right labels
1 (Practice)	32	Coca Cola	Pepsi Cola
2 (Practice)	20	bad	good
3 (Test)	20	Coca Cola-bad	Pepsi Cola-good
4 (Test)	36	Coca Cola-bad	Pepsi Cola-good
5 (Practice)	20	Pepsi Cola	Coca Cola
6 (Test)	20	Pepsi Cola-bad	Coca Cola-good
7 (Test)	36	Pepsi Cola-bad	Coca Cola-good

The DAM-4C adequately fit the data of 191 participants out of 199 (95.98%; Bonferroni-Holm method with type-I error $\alpha = .10$). A first insight into the contribution of associations to IAT responses can be obtained in a straightforward way by comparing the DAM-4C with a pure discrimination model. Such a model is obtained from the DAM-4C by constraining the eight association rates to be 0. Conversely, the two termination criteria pertaining to the association process can be left free to assume any possible value within their interval of existence. These parameters would not be interpretable because they are meaningless in a pure discrimination model. The pure discrimination model adequately fit the data of 182 participants (91.46%; Bonferroni-Holm method with type-I error $\alpha = .10$), nine less than the DAM-4C. The Akaike

information criterion (AIC) was used to compare the DAM-4C and the pure discrimination model. For 171 participants (85.93%), the AIC of the DAM-4C was lower than that of the pure discrimination model, this suggesting that automatic associations could have played some role in determining the responses of these individuals.

A total of 45 participants made no error in either the blocks *Pepsi-bad/Coca-good* or the blocks *Coca-bad/Pepsi-good*. Since the DM requires both correct and incorrect responses, it has not been estimated on the data of these participants. The comparison between the DAM-4C and the DM required us to estimate, for each participant, a third DM on the data of the practice blocks. There were 134 participants for whom all three DMs were estimable (i.e., who gave at least one incorrect response in each of the three block types). For 133 of these 134 participants, the AIC of the DAM-4C was larger than that obtained by combining the AICs of the three DMs. The following analyses were performed on the 131 participants for whom the AIC of the DAM-4C was lower than that of the pure discrimination model, and for whom the DAM-4C showed satisfactory fit.

Descriptive statistics of the estimates of DAM-4C parameters suggested the presence of aberrant estimates for the rates and the termination criteria involved in the association process. For the association rates, the mean was 21,589 to 54,209 times larger than the median. For the termination criteria K_{AC} (associations in test blocks *Pepsi-bad/Coca-good*) and K_{AI} (associations in test blocks *Coca-bad/Pepsi-good*), the mean was, respectively, 41,371 and 35,746 times larger than the median. For the correct discrimination rates and the termination criteria involved in the discrimination process, the mean was of the same order of magnitude as the median. For two incorrect discrimination rates (i.e., λ_{+-} and a_{-+}), it was twice the median (it is worth noting that there are usually only a few incorrect responses in an IAT, so that there might not be enough information to compute reliable estimates of these parameters). A total of 16.51% of the estimates of the association rates and 20.61% of the estimates of the termination criteria K_{AC} and K_{AI} were identified by Tukey's criterion as outliers and were discarded.

In line with Stefanutti et al. (2013), nine contrast measures were computed, based on the estimates of the DAM-4C. Let *c*, *p*, +, and − denote *Coca Cola*, *Pepsi Cola*, *good*, and *bad*, respectively. A *DISC* was computed for each stimulus category as the difference between the rates concerning the correct and incorrect discrimination of the category (e.g., $DISC_c = \lambda_{cc} - \lambda_{pc}$). Positive values of the *DISC*s indicate that, on average, the stimuli provide more evidence, in the time unit, about their own category than about the opposite category. An *ASSO* was computed for each stimulus category as the difference between the two association rates of the stimulus category. Positive values of $ASSO_c = \lambda_{+c} - \lambda_{-c}$ indicate that, on average, the *Coca Cola* stimuli provide more evidence, in the time unit, about the category *good* than about the category *bad*. This means that *Coca Cola* is more strongly associated with *good* than with *bad*. Similarly, positive values of $ASSO_+ = \lambda_{c+} - \lambda_{p+}$ indicate that *good* is more strongly associated with *Coca Cola* than with *Pepsi Cola*. Interpretation of $ASSO_p$ and $ASSO_-$ is opposite to that of $ASSO_c$ and $ASSO_+$. Positive values of $ASSO_p = \lambda_{-p} - \lambda_{+p}$ indicate that *Pepsi Cola* is more strongly associated with *bad* than with *good*, and positive values of $ASSO_- = \lambda_{p-} - \lambda_{c-}$ indicate that *bad* is more strongly associated with *Pepsi Cola* than with *Coca Cola*. Finally, a *DIFF* measure was also computed by contrasting the two termination criteria which pertain to the test blocks, that is $DIFF = K_{AI} - K_{AC}$. A positive value of this measure indicates that a smaller amount of evidence needs to be accumulated in the compatible blocks than in the incompatible blocks.

In line with Klauer et al. (2007), three contrast measures were computed as the difference between parameters of the two DMs which were estimated on the compatible and incompatible blocks. Let t_{0C}, a_C, and v_C be respectively the estimates of the nondecision component, speed-accuracy, and mean drift rate obtained on the compatible blocks, and t_{0I}, a_I, and v_I be the estimates of the same parameters obtained on the incompatible blocks. The three contrast measures were computed as $IAT_t = t_{0I} - t_{0C}$, $IAT_a = a_I - a_C$, and $IAT_v = v_C - v_I$. Positive values of IAT_t indicate that nondecision components require more time in the incompatible blocks than in the compatible blocks. Positive values of IAT_a indicate that speed-accuracy is more conservative in the incompatible blocks than in the compatible blocks. Finally, positive values of IAT_v indicate that the categorization task is less difficult in the compatible blocks than in the incompatible blocks.

To investigate the predictive validity of the *D* score, DAM-4C, and DM, a total of six saturated structural equation models for observed variables have been estimated. In these models, the dependent variables (brand attractiveness, taste preference, and choice of cola) were entered simultaneously, and their residuals were allowed to correlate. In models with two or more predictors, correlations between all of them were estimated. In Model 1, we established the ability of the *D* score algorithm to predict the three criteria. The total variance accounted for by the *D* score is $R^2_{\text{attractiveness}} = .04$, $R^2_{\text{taste}} = .07$, and $R^2_{\text{choice}} = .08$. In Model 2, we entered as predictors the three contrasts computed on the DM parameter estimates. The total variance accounted for by the DM is $R^2_{\text{attractiveness}} = .05$, $R^2_{\text{taste}} = .10$, $R^2_{\text{choice}} = .08$. In Model 3, we entered as predictors the nine contrasts computed on the DAM-4C parameter estimates. The total variance accounted for by the DAM-4C is higher than that accounted for by alternative scoring methods: $R^2_{\text{attractiveness}} = .06$, $R^2_{\text{taste}} = .10$, $R^2_{\text{choice}} = .12$. Table 4 presents parameter estimates of these three models. In this dataset, the DAM-4C allowed us to observe that the associations that more strongly predict cola

Table 4 Parameter estimates of three saturated structural equation models for observed variables, assessing the predictive validity of *D* score, DM, and DAM-4C in scoring the IAT

	Criterion														
	Brand attractiveness					Taste preference					Cola choice				
Model	*B*	*SE*	*β*	*p*	R^2	*B*	*SE*	*β*	*p*	R^2	*β*	*SE*	*β*	*p*	R^2
Model 1 - *D* score	0.198	0.084	**2.360**	0.018	0.04	0.268	0.088	**3.043**	0.002	0.07	0.327	0.101	**3.247**	0.001	0.08
Model 2 - DM					0.05					0.1					0.08
IAT_v	−0.016	0.021	−0.093	0.459		−0.064	0.025	**−0.312**	0.011		−0.055	0.027	**−0.252**	0.041	
IAT_a	−0.099	0.092	−0.136	0.284		−0.017	0.109	−0.019	0.877		−0.082	0.117	−0.087	0.484	
IAT_t	−0.513	0.517	−0.122	0.322		0.233	0.614	0.046	0.704		−0.024	0.659	−0.004	0.971	
Model 3 - DAM-4C					0.06					0.1					0.12
$DISC_c$	0.015	0.023	0.082	0.501		−0.056	0.024	**−0.285**	0.017		0.015	0.027	0.067	0.572	
$DISC_p$	0.039	0.024	0.181	0.109		0.048	0.026	*0.207*	0.061		0.012	0.029	0.047	0.667	
$DISC_+$	−0.022	0.023	−0.112	0.337		0.007	0.024	0.033	0.772		−0.027	0.027	−0.112	0.325	
$DISC_-$	−0.023	0.024	−0.113	0.323		0.002	0.025	0.008	0.942		−0.006	0.028	−0.026	0.817	
$ASSO_c$	0.011	0.033	0.033	0.738		0.015	0.034	0.043	0.656		−0.065	0.039	−0.160	0.092	
$ASSO_p$	−0.001	0.025	−0.004	0.965		0.024	0.026	0.085	0.364		0.063	0.030	**0.198**	0.034	
$ASSO_+$	0.021	0.040	0.054	0.592		0.024	0.041	0.058	0.558		0.127	0.047	**0.267**	0.007	
$ASSO_-$	0.020	0.028	0.073	0.460		0.012	0.029	0.039	0.683		0.004	0.033	0.011	0.905	
DIFF	−0.002	0.001	−0.171	0.115		−0.001	0.002	−0.098	0.360		−0.003	0.002	−0.169	0.110	

Note. Significant ($p < 0.05$) parameter estimates are in bold. Marginally significant ($p < 0.10$) parameter estimates are in italics

choice are $ASSO_p$ and $ASSO_-$. Controlling for the *D* score algorithm, the incremental validity of the DM (Model 4) is $\Delta R^2_{\text{attractiveness}} = .04$ ($p = .84$), $\Delta R^2_{\text{taste}} = .03$ ($p = .86$), $\Delta R^2_{\text{choice}} = .03$ ($p = .86$), whereas the incremental validity of the DAM-4C over the *D* score (Model 5) is $\Delta R^2_{\text{attractiveness}} = .06$ ($p = .81$), $\Delta R^2_{\text{taste}} = .07$ ($p = .79$), $\Delta R^2_{\text{choice}} = .09$ ($p = .76$). Controlling for both *D* score and DM, the incremental validity of the DAM-4C (Model 6) is $\Delta R^2_{\text{attractiveness}} = .04$ ($p = .84$), $\Delta R^2_{\text{taste}} = .07$ ($p = .79$), $\Delta R^2_{\text{choice}} = .08$ ($p = .77$).

Study 2: Validation of DAM-4C parameters

The present study aims at validating the parameters of the DAM-4C. A known-groups method is used to compare parameter estimates obtained on white and black respondents to a Black-White IAT. Such an IAT measures the implicit preference for white individuals over black individuals.

In general, we expect to find implicit ingroup favoritism (Tajfel & Turner, 1979) in both cultural groups. Let *w*, *b*, +, and − denote *White People*, *Black People*, *good*, and *bad*, respectively. Operatively, we expect that λ_{+w}, λ_{w+}, λ_{-b}, and λ_{b-} will be higher in white people and that λ_{+b}, λ_{b+}, λ_{-w}, and λ_{w-} will be higher in black people. Previous research (Anselmi et al., 2011) found evidence that the contribution of positive associations to the overall IAT effect is stronger than that of negative associations. Hence, we expect white people to show higher average estimates of λ_{w+} and λ_{+w} compared with λ_{b-} and λ_{-b} and black people to show higher average estimates of λ_{b+} and λ_{+b} compared with λ_{-w} and λ_{w-}. Also, we expect to replicate previous research that observed (1) higher ingroup favoritism in white people than in black people (Greenwald et al., 1998; Jost et al., 2004; Nosek et al., 2002; Nosek et al., 2007b) and (2) outgroup favoritism in black people when bad word stimuli are made more salient in the IAT procedure than good words (Axt et al., 2018). Previous results showing lower ingroup favoritism in black people compared with white people have been interpreted in light of system justification theory (Jost & Banaji, 1994), according to which people with lower social status, in order to meet their need to view the world as fair and preserve the status quo, may implicitly retain certain cultural values and stereotypes that associates black people with bad. Hence, we expect that:

1. the differences between (*a*) λ_{w+} and λ_{w-}, (*b*) λ_{+w} and λ_{-w}, (*c*) λ_{b+} and λ_{b-}, and (*d*) λ_{-b} and λ_{+b} will be larger in white people than in black people;
2. λ_{b-} and λ_{-b} in black people will be higher than λ_{w-} and λ_{-w} in white people.

Lastly, we expect to find no differences in correct and incorrect discrimination rates across groups, nor in the termination criteria.

The data consist of the responses to the Black-White IAT available at

https://implicit.harvard.edu/implicit. Of the participants who completed the IAT from 30 July 2018 to 8 August 2018, 355 self-declared as white (214 females) and 36 as black (27 females).

Participants were presented with the Black-White IAT according to the structure in Table 5. A total of 24 pictures were used for representing the target categories *White People* and *Black People*, and 16 words were used for representing the attribute categories *good* (glorious, happy, joy, laughter, love, peace, pleasure, wonderful) and *bad* (agony, awful, evil, failure, horrible, hurt, nasty, terrible). The stimuli were presented in the center of the computer screen in an alternating fashion, and participants were asked to categorize them by pressing, as quickly and accurately as possible, the response key "E" or "I". A red "X" appeared in case of a mistake, and it disappeared after the correct response was given.

The DAM-4C fit the data of 342 participants (87.47%; Bonferroni-Holm method with type-I error α = .10). Conversely, the pure discrimination model (see Sect. 6) fit the data of 284 participants (72.63%), 58 less than the DAM-4C. For 346 participants (88.49%), the AIC of the DAM-4C was lower than that of the pure discrimination model. This result suggests that there could be non-negligible associations in the responses of these individuals. The following analyses were performed on the 298 participants (272 whites, 26 blacks) for whom the AIC of the DAM-4C was lower than that of the pure discrimination model, and for whom the DAM-4C showed satisfactory fit.

Descriptive statistics of the estimates of DAM-4C parameters suggested the presence of aberrant estimates for the rates and the termination criteria involved in the association process. For the association rates, the mean was 13,467 to 83,677 times larger than the median. For the termination criteria K_{AC} (associations in test blocks *Black-bad/White-good*) and K_{AI} (associations in test blocks *White-bad/Black-good*), the mean was, respectively, 39,669 and 17,994 times larger than the median. For the correct discrimination rates and the termination criteria involved in the discrimination process, the mean was of the same order of magnitude as the median, whereas for the incorrect discrimination rates it was from two to three times the median (it is worth noting that, typically, there are only a few incorrect responses in an IAT). A total of 15.65% of the estimates of the association rates, and 19.30% of the estimates of termination criteria K_{AC} and K_{AI} were identified by Tukey's criterion as outliers and discarded.

Table 5 Structure of the Black-White IAT

Block type	No. of trials	Left labels	Right labels
1 (Practice)	20	Black People	White People
2 (Practice)	20	bad	good
3 (Test)	20	Black People-bad	White People-good
4 (Test)	52	Black People-bad	White People-good
5 (Practice)	28	White People	Black People
6 (Test)	20	White People-bad	Black People-good
7 (Test)	52	White People-bad	Black People-good

Figure 8 displays mean parameter estimates (and 95% confidence intervals) for white (white bars) and black (gray bars) respondents. In the figure, *w*, *b*, +, and − in the λ parameters denote *White People*, *Black People*, *good*, and *bad*, respectively. K_{AC} and K_{AI} denote the termination criteria of the associations involved in test blocks *Black-bad/White-good* and *White-bad/Black-good*, respectively.

Between-group differences in correct discrimination rates are small or negligible (average Cohen's d = − .16). Black people showed higher incorrect discrimination rates (average Cohen's d = − .36), especially when they were asked to categorize faces of black people (Cohen's d = − .64). We do not have an interpretation for this unexpected result, and we think that it would be premature to reach any kind of conclusion. Yet, a possible interpretation is that black people perceive less of a difference between faces of white and black people.

With regard to association rates, estimates of λ_{w+}, λ_{-b}, and λ_{b-} are higher in white people (Cohen's ds =.15, .36, and .27

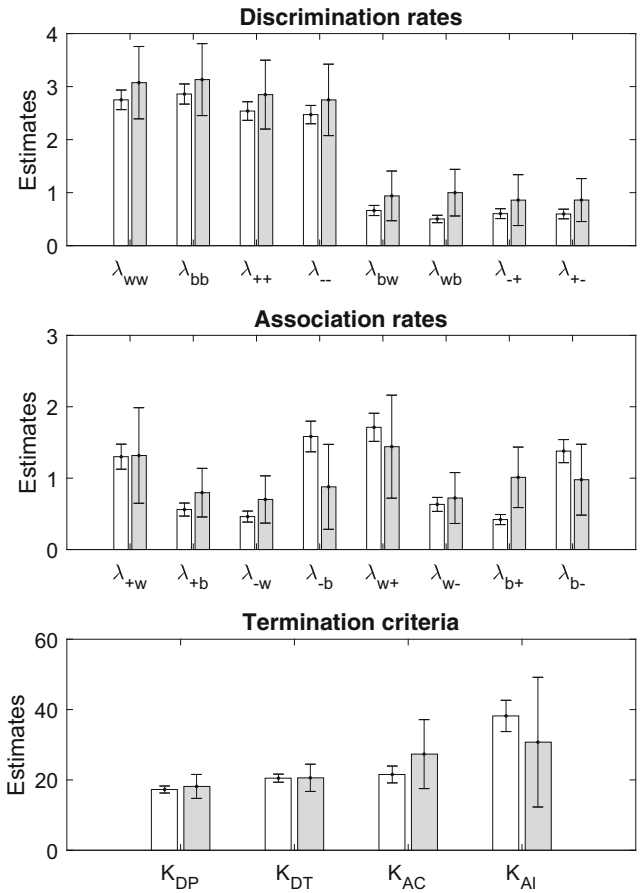

Fig. 8 Mean DAM-4C parameter estimates (and 95% confidence intervals) for white (white bars) and black (gray bars) respondents.

for λ_{w+}, λ_{-b}, and λ_{b-}, respectively) and λ_{+b}, λ_{b+}, and λ_{-w} are higher in black people (Cohen's *d*s = − .27, − .84, and −.33 for λ_{+b}, λ_{b+}, and λ_{-w}, respectively), as we expected. We interpret these results as supporting the validity of these parameters. Contrary to our expectations, λ_{+w} and λ_{w-} are approximately equal across groups (average Cohen's d = − .05). Further research is needed to interpret this result, but it is evident that, at least in this sample, these two components do not influence the overall IAT effect, and might not be critical in determining differences across black and white people in the phenomenon of ingroup favoritism.

With regard to our expectation that the association parameters of the DAM-4C would have detected the positive association primacy effect, we observed that (*a*) white people did not show higher average estimates of λ_{w+} and λ_{+w} compared with λ_{b-} and λ_{-b} (Cohen's *d*s =.16, . 10 for λ_{w+} vs. λ_{b-} and λ_{+w} vs. λ_{-b}, respectively), and (*b*) black people did not show higher average estimates of λ_{+b} compared with λ_{-w} (Cohen's d = .13) and λ_{w-} (Cohen's d = − .19). In black people, the estimates of λ_{b+} are higher than those of λ_{w-} (Cohen's d = .37) and λ_{-w} (Cohen's d = .26). This result was expected and replicates previous research showing that a positive evaluation of the ingroup, rather than a negative evaluation of the outgroup, is at the basis of ingroup favoritism. This result suggests that the positive association primacy effect may be due only to a single component of the overall IAT effect, which is the association between good stimuli and black labels.

The differences between (*a*) λ_{w+} and λ_{w-}, (*b*) λ_{+w} and λ_{-w}, (*c*) λ_{b+} and λ_{b-}, and (*d*) λ_{-b} and λ_{+b} are larger in white people than in black people (Cohen's *d*s =.27, .08, − .70, and .49, respectively). Differences across groups are especially large for λ_{b+} and λ_{b-} and for λ_{-b} and λ_{+b}, supporting our predictions and adding some nuances to the interpretation of the known effect that implicit ingroup favoritism is larger in white people than in black people (Greenwald et al., 1998; Jost et al., 2004; Nosek et al., 2002; Nosek et al., 2007b). With regard to the outgroup favoritism observed in black people when bad words are made more salient than good words stimuli in the IAT procedure (Axt et al., 2018), we found that λ_{b-} and λ_{-b} in black people were higher than λ_{w-} and da_{-w} in white people (Cohen's d = .37 for λ_{b-} in blacks vs. λ_{w-} in whites, and .53 for λ_{-b} in blacks vs. λ_{-w} in whites).

Turning to termination criteria, we observed no differences across groups in the termination criteria pertaining to discrimination. For both white and black respondents, the double categorization blocks were more difficult than the single categorization blocks. For white respondents, the amount of evidence to be collected in their incompatible block (*White-bad/Black-good*) was larger than the amount of evidence to be collected in their compatible block (*White-good/Black-bad* condition). In black respondents, both critical tasks were equally difficult.

Discussion

A new formulation of the DAM has been presented in which the processes involved in stimuli discrimination and automatic association remain separate and independent instead of being collapsed into a single process. Results of theoretical and simulation studies suggest that the DAM-4C outperforms the DAM. The IAT effect is found to vary with the association rates of the DAM-4C and not with those of the DAM. The DAM-4C allowed us to observe that the choice of Coca Cola over Pepsi Cola is mostly due to two specific associations: (*a*) Pepsi Cola is judged as more negative than positive, and (*b*) negative attributes are associated more strongly with Pepsi Cola than with Coca Cola. In addition, the association rates estimated on data from a Black-White IAT are in line with expectations. Providing information about an association mechanism is a peculiar feature of the DAM-4C, which enables a fine-grained decomposition of the IAT effect.

Compared with the DAM, the DAM-4C leaves some degrees of freedom in the way of modeling the evidence accumulation process. In the present work, the assumption was made that the two categories involved in the same process (which could be either the discrimination process D or the association process A) accumulate evidence in parallel. The choice of considering the PRM is consistent with such an assumption. Alternatively, the assumption could have been made that the two categories accumulate evidence serially. In this case, the diffusion model (Ratcliff, 1978; Ratcliff & Rouder, 1998) would have been an option for modeling evidence accumulation. It is worth noting that, regardless of the assumption of parallel or serial processing, in the DAM-4C, discrimination D and association A remain two parallel and independent processes. Conversely, because of the sum between the two processes associated to the same response key, in the DAM there is no separation between discrimination and association. Moreover, whereas the sum of two Poisson processes results in another Poisson process, this is not the case for the sum of two diffusion processes. This prevents us from modeling evidence accumulation in the DAM via a diffusion model.

A limit of the present study is that the number of trials in the two IATs was not sufficient for obtaining reliable estimates of all parameters of the DAM-4C. For some participants, large values were observed for the rates and termination criteria involved in the association process. The goodness-of-recovery study suggests that the responses given by these participants might have been more often determined by a discrimination process than by an association process. A very simple way to test whether associations are negligible requires one to compare the DAM-4C with a pure discrimination model. This was done in the present study. However, it is worth noting that the DAM-4C and the pure discrimination model represent two extremes of a continuum. Between these two

extremes, several variants of the DAM-4C can be specified that differ in the number and type of associations involved in the responses to the IAT. These models can provide useful information about the specific associations of the individuals.

Open Practices Statement The codes for estimating and testing the DAM-4C, as well as the data, are available at: https://osf.io/fg8ht/

References

Aldous, D. and Shepp, L. (1987). The least variable phase type distribution is Erlang. *Stochastic Models*, 3:467–473.

Anselmi, P., Vianello, M., and Robusto, E. (2011). Positive associations primacy in the IAT: A many-facet Rasch measurement analysis. *Experimental Psychology*, 58(5):376–384.

Anselmi, P., Vianello, M., Stefanutti, L., and Robusto, E. (2013). A Poisson race model for the analysis of the Implicit Association Test. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, 20(3):249–261.

Axt, J., Moran, T., and Bar-Anan, Y. (2018). Simultaneous ingroup and outgroup favoritism in implicit social cognition. *Journal of Experimental Social Psychology*, 79:275–289.

Buchholz, P., Kriege, J., and Dohndorf, I. (2014). *Input modeling with phase-type distributions and Markov models. Theory and applications*. Springer, New York, NY.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., and 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(4):469–487.

Ephraim, Y. and Mark, B. L. (2012). Bivariate Markov processes and their estimation. *Foundations and Trends in Signal Processing*, 6(1):1–95.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.

Greenwald, A. G., Nosek, B. A., and Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2):197–216.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors (1983). *Understanding robust and exploratory data analysis*. John Wiley & Sons, New York, NY.

Jost, J. T. and Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology*, 33(1):1–27.

Jost, J. T., Banaji, M. R., and Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, 25(6):881–919.

Klauer, K. C., Voss, A., Schmitz, F., and Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology*, 93(3):353–368.

Lane, K. A., Banaji, M. R., Nosek, B. A., and Greenwald, A. G. (2007). Understanding and using the Implicit Association Test: IV What we know (so far) about the method. In Wittenbrink, B. and Schwarz, N., editors, *Implicit Measures of Attitudes*, pages 59–102. The Guilford Press, New York, NY.

Luce, R. D. (1996). The ongoing dialog between empirical science and measurement theory. *Journal of Mathematical Psychology*, 40:78–98.

Meissner, F. and 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(1):45–69.

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101–115.

Nosek, B. A., Greenwald, A. G., and Banaji, M. R. (2007a). The Implicit Association Test at age 7: A methodological and conceptual review. In Bargh, J. A., editor, *Automatic Processes in Social Thinking and Behavior*, pages 265–292. Psychology Press, New York, NY.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., and Banaji, M. R. (2007b). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18(1):36–88.

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

Ratcliff, R. and Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, 9:347–356.

Ratcliff, R. and Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, 9:438–481.

Stefanutti, L., Robusto, E., Vianello, M., and Anselmi, P. (2013). A discrimination-association model for decomposing component processes of the Implicit Association Test. *Behavior Research Methods*, 45:393–404.

Stefanutti, L., Vianello, M., Anselmi, P., and Robusto, E. (2014). GRace: A MATLAB-Based application for fitting the discrimination-association model. *The Spanish Journal of Psychology*, 17:e73.

Townsend, J. T. and Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge University Press, New York, NY.

Voss, A. and Voss, J. (2007). Fast-dm: a free program for efficient diffusion model analysis. *Behavior Research Methods*, 39:767–775.

Voss, A. and Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology*, 52:1–9.