Title
Modeling Underlying Mechanisms of the Implicit Association Test

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 33(33)

ISSN
1069-7977

Authors
Quek, Boon-Kiat
Ortony, Andrew

Publication Date
2011

Peer reviewed

Modeling Underlying Mechanisms of the Implicit Association Test

Boon-Kiat Quek (boonkiat.quek@northwestern.edu)
Andrew Ortony (ortony@northwestern.edu)
Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore, and
Northwestern University, Evanston, IL 60208, USA

Abstract

The Implicit Association Test (IAT) is designed to measure implicit attitudes, and is often claimed to reveal prejudicial attitudes that are at odds with explicit attitudes. Numerous proposals as to the information processing mechanisms underlying human performance on the IAT have been suggested, many or all of which may well play a role. This makes it difficult to study them and their interactions experimentally in an efficient manner. We describe a localist connectionist model that simulates human performance on the IAT and that allows us to explore many of the proposed explanations, by comparing the results with observations from actual experiments with human subjects. By simulating the performance of virtual subjects, the model also makes it possible to conduct "theoretical" experiments that could not be undertaken with real subjects in the real world.

Keywords: Associations; attitudes; simulation; localist connectionist networks.

Introduction

The Implicit Association Test (IAT; Greenwald, McGhee & Schwartz, 1998) is a computer-administered test (see e.g., https://implicit.harvard.edu/implicit/) designed to measure automatic associations between concepts, where such associations are assumed to underlie implicit attitudes towards attitude objects (Greenwald & Banaji, 1995). It has been used in a variety of domains, including age, race, self-esteem, sexual orientation, and so on. By being implicit, these attitudes are assumed to be beyond conscious control (Greenwald & Banaji, 1995) and thus cannot be assessed with explicit measures such as self-reports. Instead, the IAT seeks to measure implicit attitudes by recording subjects' response times as they complete a number of tasks, the performance on which is assumed to involve accessing automatic associations. Such tasks involve pressing a *Left* or *Right* key on a keyboard to classify sequentially presented stimuli into one of two categories. Each category comprises one of two target concepts (e.g., *flower* or *insect*) crossed with one of two evaluative attributes (e.g., *pleasant* or *unpleasant*). This results in four possible pairings, two of which are, by hypothesis, incompatible (e.g. *insect* and *pleasant*) and two of which are compatible.

A central assumption of the IAT is that the stronger the associations between an attitude object and its evaluative attributes, the stronger the implicit attitude. Thus, a person with a negative attitude towards insects is assumed to have a correspondingly strong association between the concept *insect* and some general concept for negativity. The idea is that the strength of these associations will be reflected in people's response times on the classification tasks, with longer latencies for incompatible pairings than for compatible pairings. This difference in response times between trials for incompatible and compatible pairings is known as the IAT effect (Greenwald, McGhee & Schwartz, 1998).

Although the IAT is widely used (Greenwald, Poehlman, Uhlmann & Banaji, 2009), its history is rife with controversy. Questions have been raised as to whether measuring automatic associations is a valid way of measuring implicit attitudes (Blanton & Jaccard, 2006), and whether it is indeed measuring automatic associations at all (De Houwer, 2001; Brendl, Markman & Messner, 2001). Researchers have also wondered whether IAT effects might be partially due to other factors such as differences in stimulus familiarity (Dasgupta et al, 2000; Ottaway, Hayden & Oakes, 2001) or salience asymmetries between target concepts (Rothermund & Wentura, 2004), and whether they might merely be a reflection of prevailing cultural norms rather than of genuine attitudes (Karpinski & Hilton, 2001).

Given these various proposals, several accounts of the processing mechanisms underlying the IAT have been advanced. These include the *figure-ground asymmetry model* (Rothermund & Wentura, 2001), a *random-walk model* (Brendl, Markman & Messner, 2000), a *diffusion model* (Klauer, Voss, Schmitz & Teige-Mocigemba, 2007), an *acquired meaning of response keys* account (De Houwer, 2001) and a suggestion that longer response latencies are the result of *task-set switching* overheads (Mierke & Klauer, 2001; Klauer & Mierke, 2005) occasioned by moving between compatible and incompatible task blocks. Finally, a *quad model of implicit task performance* (Conrey, Sherman, Gawronski, Hugenberg & Groom, 2005) postulates that multiple processes are at work – some automatic and others more controlled (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977).

These differing explanations are not necessarily mutually exclusive. It might be that some, or even all of them, play a role, in which case one would need a model that could, among other things, account for (i) the interaction between automatic and control processes (e.g., Mierke & Klauer, 2001; Herd, Banich & O'Reilly, 2006), (ii) cue and response competition induced by the interference of competing stimuli in incompatible trials (e.g., Praamstra & Seiss, 2005), (iii) resolution of these conflicts (e.g., through the mutual inhibition of competing action tendencies, Atkinson & Birch, 1970), and (iv) increase in cognitive processing (e.g., Snyder-Tapp & Dale, 2009) and reduction in speed and accuracy (Klauer & Mierke, 2005) on incompatible trials.

To investigate some of these issues, we developed a computational model intended to emulate human performance on the IAT. The model handles the generation of output behavior, and allows us to investigate such issues as re-

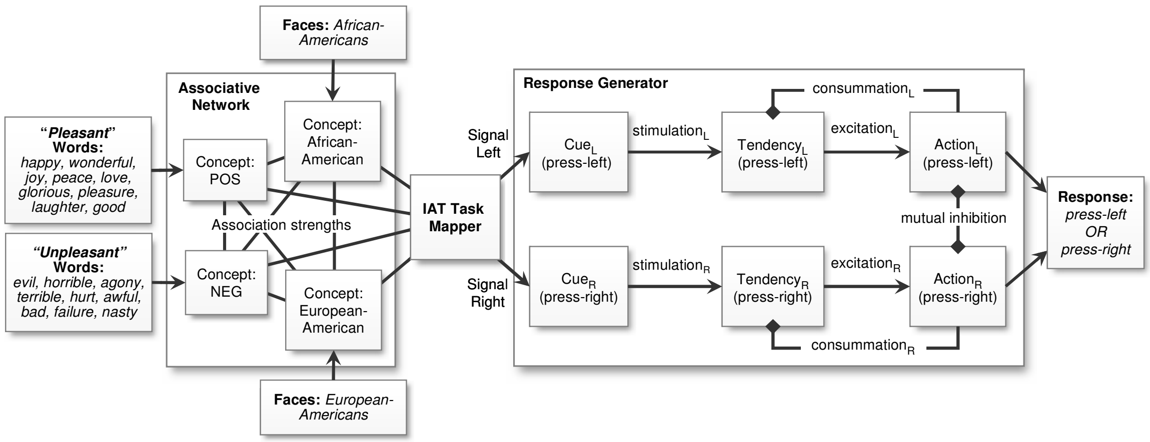

Figure 1. Virtual subject model for simulation of Implicit Association Tests

sponse competition and conflict resolution (see (ii) and (iii) above). The work we present here focuses on the application of the model to the Race-IAT because it is the most controversial and is the version of the test for which there exists the most empirical data (e.g., Greenwald et al., 1998; Karpinski & Hilton, 2001) – this latter fact giving us more opportunity to compare human and model performance.

Model Overview

The model employs a spreading activation algorithm operating on a localist connectionist network (see, e.g., Page, 2000). It emulates the multiple pathways leading from the initial perception of a visual stimulus (in our case, a word or image) to the automatic activation of concepts in memory through to the motor response called for by the particular task instructions. As in all localist connectionist models, the nodes in the network are semantically meaningful entities representing concepts or processes, with connections between them representing associative strengths. The actual behavior of such networks depends critically on the details of their topology and the algorithms used. Mathematically speaking, the network can be defined as a graph $G=\{V, E\}$ with nodes V and connections or edges, E. Each node v_i contains a label $name_i$ representing a particular concept or process, and a quantity x_i representing its activation level:

$$v_i=<name_i\,, x_i>;\ \ v_i\in V\,;\, x_i\in[-1, 1]. \tag{1}$$

Connections or edges take the form:

$$\varepsilon_{i,j}=<v_i\,, v_j\,, w_{i,j}>;\ \ \varepsilon_{i,j}\in E\,;\, v_i\,, v_j\in V\,;\, w_{i,j}\in[-1, 1]\,, \tag{2}$$

where $w_{i,j}$ is the associative strength between nodes v_i and v_j. The propagation rule for successive iterations is:

$$x_i(k+1) = (1-\delta)x_i(k) + \alpha \sum_{\varepsilon_{j,i}\in E} x_j(k)\cdot w_{j,i}(k)\,, \tag{3}$$

where α is the gain (set to 0.2) and δ is a decay parameter (set to 0.001) that reduces activation over time. In this way, activation spreads to v_i from each neighbor v_j at a rate proportional to $w_{j,i}$ in each time step. Parameter values in these ranges ensure that activation levels do not saturate prematurely and have sufficient momentum to accumulate. Prior to the introduction of a perturbation factor (see below), the connections weights are initialized with magnitudes of 0.5.

If as is widely believed, the IAT effect is indeed the result of patterns of activation of underlying associations between target concepts and evaluative attributes, then the kind of network model we are discussing is ideal for simulating the behavior of human subjects in the IAT; it simply involves configuring the network so that it represents the relevant associative biases in virtual subjects.

Modeling Virtual Subjects In addition to the *Associative Network*, each simulated subject has two other main components: a *Task Mapper*, and a *Response Generator* (see Figure 1). The *Task Mapper* dynamically transmits activation levels of concepts from the *Associative Network* to the *press-left* or *press-right* nodes, which serve as cues for the corresponding actions or motor responses. The *Response Generator*, when necessary, resolves competition between these cues to generate one of the two motor responses (i.e., key presses). When simulating the performance of a group of subjects, each virtual subject is configured with the same network topology, but with randomly distributed connection weights so as to produce a varied sample of simulated subjects. This is achieved by perturbing connection weights with normally distributed noise of mean $\mu_{j,i} = 0$ and standard deviation $\sigma_{j,I} = 0.1$, which is large for $w_{j,i}\in[-1, 1]$. The result is that, with a default $w_{j,i}$ of ±0.5, the majority (95%) of connections across subject populations ends up with magnitudes in the [0.3, 0.7] range. Introducing random perturbations to connection weights further ensures that the outcome of the task is not critically dependent on any particular distribution of connection weights.

Mapping Stimuli to Concepts for the Race-IAT The main concepts for the Race-IAT are the target concepts AFRICAN-AMERICAN (AA) and EUROPEAN-AMERICAN (EA), along with the generic evaluative attributes POS-itive and NEG-ative. Whereas there are many variants of the Race-IAT, in the version we are modeling, the verbal (input) stimuli for the Race-IAT are words belonging to the semantic fields *pleasant* (e.g., *happy, wonderful, joy*) or *unpleasant* (e.g., *evil, horrible, hurt*). The pictorial stimuli are faces of (all presumed to be unfamiliar) European-American or African-

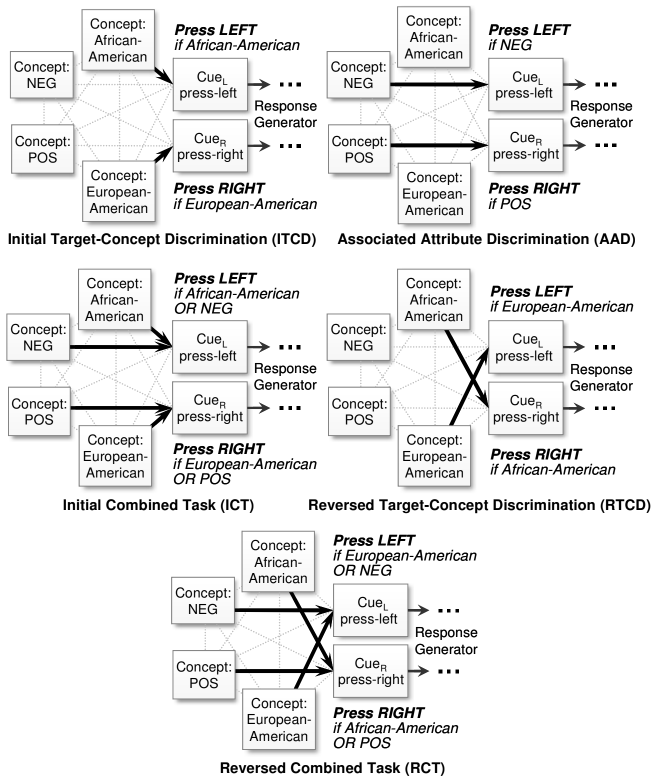

Figure 2. Mapping from concepts to cues for each task block

American individuals. As each input is presented to a virtual subject, a signal indicating the occurrence of a specific verbal or pictorial stimulus is transmitted to the subject's network. For instance, if the word *wonderful* is presented, the concept WONDERFUL is activated, which in turn increases the activation level of the concept POS. If a picture of a random European-American face is presented, it activates a concept for an individual European-American face, which in turn activates the concept EUROPEAN-AMERICAN. In both cases, the path lengths from input nodes to EA or AA, and to POS or NEG are the same.

Task Switching and Mapping Concepts to Cues for Motor Responses The *Task Mapper* serves as a network switchboard that, depending on the task, transmits the accumulated activation from target concepts and their evaluative attributes to the concepts corresponding to *Left* and *Right* key-press responses. These latter concepts need to be distinguished from the actual key-press responses themselves, for which reason we refer to them as response *cues*. For instance, if the current task were a *Right* key press for *pleasant* words or pictures of European-Americans, the *Task Mapper* would establish connections from both POS and EUROPEAN-AMERICAN to cue_R (see bottom half of the ICT panel in Figure 2). The assigned connections remain enforced for each block of trials, but are reconfigured for each subsequent task. To some extent, this reconfiguration process is an approximation of task-set switching (Klauer & Mierke, 2005) and emulates the dynamic remapping of valence to keys across different task blocks (De Houwer, 2001), although a fuller account (as explained in the Discussion section below) might require additional mechanisms.

Response Competition and Mutual Inhibition The *Response Generator* is a network-based implementation of the Cue-Tendency-Action (CTA) model (Revelle, 1986), which is a re-parameterization of Atkinson and Birch's (1970) Dynamics of Action theory. CTA models the dynamic interaction between conflicting tendencies and conflicting actions. The reduction in action-tendencies that results from the successful completion of the corresponding action is a form of negative feedback referred to in Figure 1 as consummation. In the present context this captures the fact that pressing a key satisfies the need to produce a response. Finally, mutual or cross inhibition between two competing actions (i.e., pressing *Left* vs. *Right*) ensures that only one of the two actions will be performed.

Performing the Task The result of the interaction of the processes and representations described above is a competition between all propagation pathways from stimulus input nodes to the final key-press output nodes. For instance, suppose the task is to press a *Left* key for pictures of European-American faces or pleasant words and a *Right* key for pictures of African-American faces or unpleasant words. When a European-American face is presented to the model, the concept of EUROPEAN-AMERICAN is eventually activated, and relatively more activation is transmitted to cue_L in the *Response Generator*. However, if the simulated subject is initially configured with a stronger connection between EUROPEAN-AMERICAN and NEG, activation is also transmitted to cue_R, leading to competition with cue_L. This results in competition between the *Left* and *Right* responses, reducing the degree to which activation accumulates in the *Left* response node, and as a consequence more time is required for it to reach the threshold for the actual response.

Simulation Results

A schematic diagram of the simulation system is shown in Figure 3. Twenty-five simulated subjects (each having its own unique associative network) were assigned to each of four groups differing in their associative configurations as follows: (a) excitatory associations between AFRICAN-AMERICA and NEG, EUROPEAN-AMERICAN and POS, and inhibitory, otherwise; (b) excitatory associations between EUROPEAN-AMERICAN and NEG, AFRICAN-AMERICAN and POS, and inhibitory, otherwise; (c) equal associations (modulo random perturbations described above) between both target concepts and evaluative attributes; and (d) no associations (i.e., connection weights with a mean of zero) between target concepts and evaluative attributes. Each virtual subject was put through all five standard IAT tasks (see Figure 2), namely, the initial target concept discrimination (ITCD), associated attribute discrimination (AAD), initial combined task (ICT), reversed target concept discrimination (RTCD), and reversed combined task (RCT). On each trial, the subject was presented with a simulated verbal or pictorial

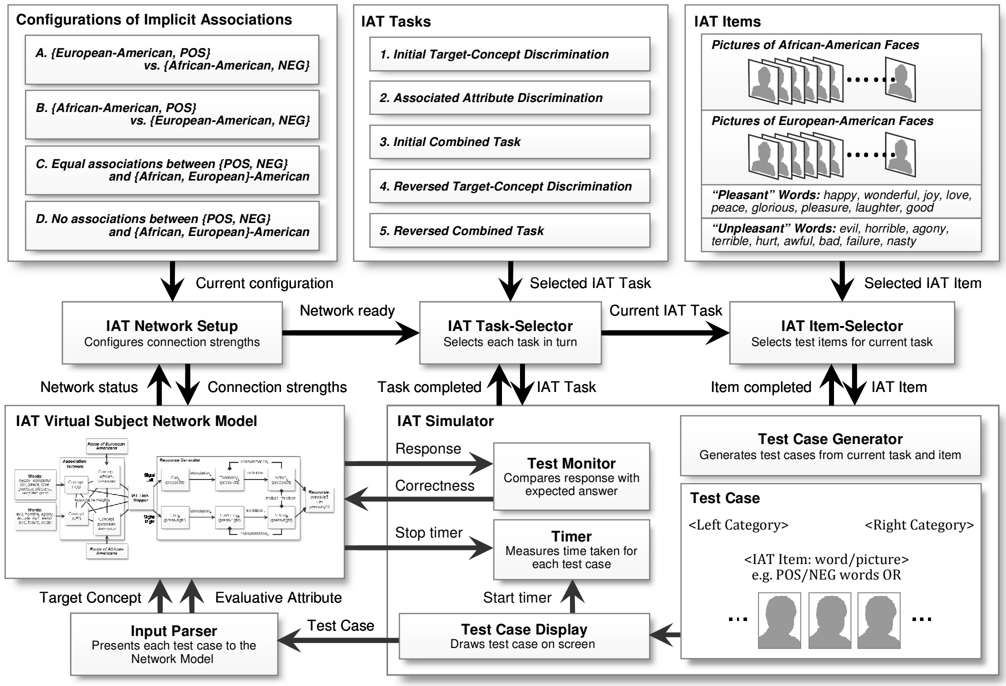

Figure 3. Schematic of the IAT simulation system

stimulus and the number of iterations (typically between 20 and 50) taken to produce a response was recorded. This was then transformed into a simulated response time (in milliseconds) by a scaling factor calibrated to the specifications of the computer on which the simulation was run. In this way, mean response times across subjects and conditions were scaled to the same order of magnitude as those reported in experiments with human subjects. The IAT effect is then taken to be the difference between mean response times in the ICT and RCT blocks. As in actual IATs, only correct responses are analyzed.

Plots of response times in Figure 4 for all task blocks and *magnitudes* of the IAT effect in each group indicate, as expected, the presence of IAT effects in the experimental groups (a, b). The effect in (a) is a preferential evaluation of European-American stimuli ($t_a(24)=22.3$, $p_a<.001$), while African-American stimuli are preferentially evaluated in (b) ($t_b(24)=-18.0$, $p_b<.001$). No significant effects are found in the control groups (c, d) ($t_c(24)=-0.923$, $p_c=0.365$; $t_d(24)=-1.22$, $p_d=0.235$). Results of planned comparisons of IAT effects between the experimental (a, b) and control (c, d) conditions are shown in Table 1.

Just as in actual experiments with human subjects (e.g., Greenwald et al., 1998; Klauer et al., 2007), response latencies on incompatible-pairings tasks are longer compared to those for other tasks, while latencies on compatible-parings tasks are similar to those on non-combined task blocks (i.e., ITCD, AAD, and RTCD). To ensure replicability, the simulation was run ten times, and as can be seen from Table 2, when the results are pooled across all ten simulations, the pattern of results is the same.

Discussion

Propagation of activation from input nodes representing verbal or pictorial stimuli to output nodes representing key-press responses occurs along multiple pathways. The pattern of propagation is determined by the configurations of associations among target concepts and evaluative attributes, as well as by the task mapping from these concepts to the response cues. Mutual reinforcement of compatible pathways and inhibition between incompatible pathways both lead to patterns of responses (including IAT effects) similar to those observed in laboratory experiments.

One criticism of our model is that the connection weights among concepts are arbitrary rather than empirically grounded. With respect to the sign on these weights, we are on safe ground because the only cases that could be in contention are the ones we manipulate in our four configurations. With respect to the magnitudes, one response would be that while weights varied randomly within a reasonably wide range, all subjects in the experimental conditions nevertheless showed IAT effects, suggesting that the actual values are probably not critical. This might appear counter-intuitive, but in fact it is consistent with data from the real world indicating that thousands of people from all walks of life (which can be thought of as an analog of randomly distributed weights) routinely show IAT effects (Nosek, Banaji & Greenwald, 2002).

One might be tempted to interpret the IAT effect as evidence for an implicit negative attitude towards the less preferred target concept. However, Greenwald et al. (1998) carefully avoided any such claim, instead, always discussing

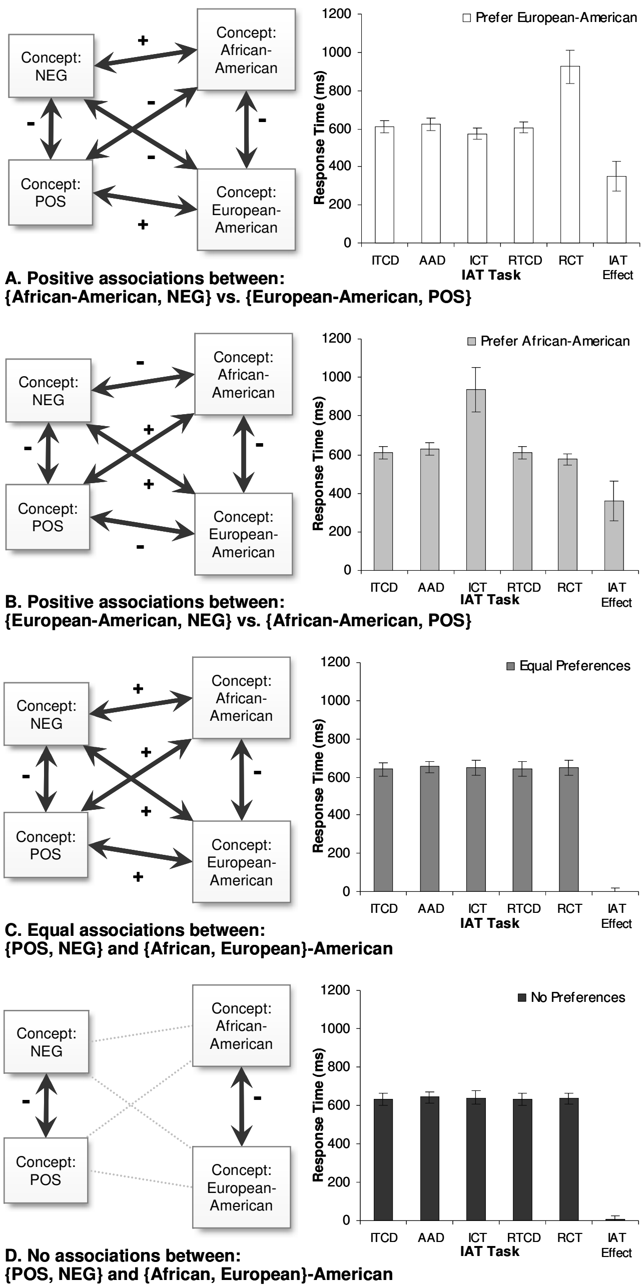

Figure 4. Simulation of four configurations between target and evaluative attribute concepts. Left panels: directions of associative strengths. Right panels: mean response times for each task, and *magnitudes* of IAT effects. Error bars: standard deviations.

the IAT effect in terms of a relative preference of one target concept over the other, an issue explicitly raised by Brendl, Markman, and Messner, (2001). In our terms, what really matters is the relative differences in weights between target concepts and their evaluative attributes. This being the case, if the network were configured with no connections to NEG, we would predict that an IAT effect might still be apparent, provided that the connections to POS of the two target concepts differed sufficiently. In fact, we tested this possibility by simulating two new groups, namely, (e) a group with strong positive associations (mean 0.8) between EUROPEAN-

Table 1: Planned comparisons between test and control groups

Control	A. Prefer EA	B. Prefer AA
C. Equal associations	t=22.0, p<.001	t=-17.5, p<.001
D. No associations	t=22.0, p<.001	t=-17.6, p<.001

Table 2: Replicating IAT Effects over 10 simulations

Condition	IAT Effect (ms) over 10 simulations		
	M	*SD*	*95% CI*
A. Prefer EA	360.05	22.28	[346.24, 373.86]
B. Prefer AA	-365.08	15.10	[-374.44, -355.73]
C. Equal associations	-0.18	6.44	[-4.18, 3.81]
D. No associations	0.17	2.94	[-1.65, 1.99]

AMERICAN and POS, and weak positive associations (mean 0.2) between AFRICAN-AMERICAN and POS, and (f), the converse of (e). Consistent with the idea that the presence of an IAT effect does not require the presence of a negative or prejudicial attitude, IAT effects emerged despite there being no connections to NEG, as shown in Figure 5. Clearly, it would be impossible to test such a prediction using standard experimental procedures because there would be no way of knowing the positivity or negativity of subjects' implicit associations. We think that the fact that one can test through simulation a prediction about implicit attitudes that could not be tested empirically in the real world is an appealing aspect of our approach.

So far, we have only focused on modeling automatic as opposed to controlled processes. Therefore, issues related to presumed discrepancies between implicit and explicit measures of attitudes, the role of selective attention (Fazio, 2001) or self-presentation bias (Dasgupta et al., 2000) have not been explored. Our intuition is that a second network operating on top of the first might make it possible to model these more cognitively elaborated phenomena. Similarly,

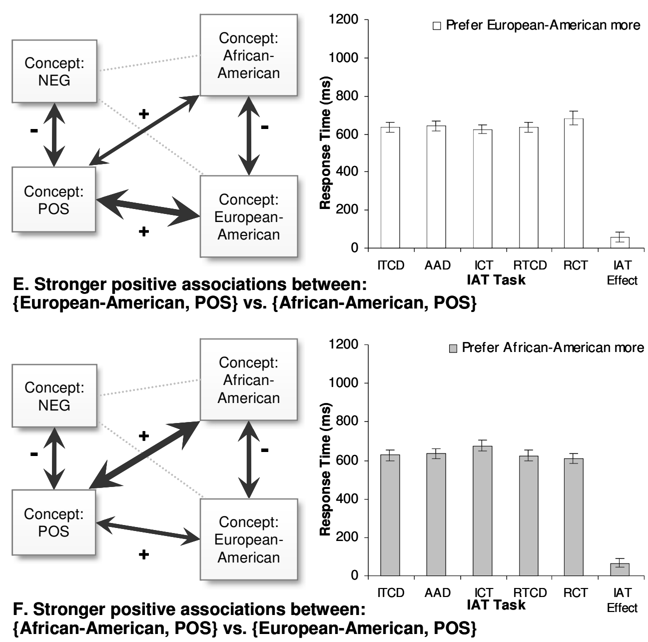

Figure 5. Simulations showing presence of IAT effects without associations with NEG. Arrow thickness: relative associative strengths. Error bars: standard deviations.

more work needs to be done to account for any effects on response latencies that might result from task-switching overheads (Klauer & Mierke, 2005) or from the dynamic remapping of valence to the left and right keys (De Houwer, 2001). These issues might be addressed by augmenting the network with nodes representing IAT tasks that would modify connection weights between attitude objects and the *Left* or *Right* response cues, instead of having the *Task Mapper* assign these weights before each task.

Our work on the IAT effect is in the same spirit as models of the Stroop effect (Cohen, Dunbar & McClelland, 1990) and of various lexical effects in language processing (Ortony & Radin, 1989). Such models employ simple algorithms running on simple networks of associations between representations in memory, but have the advantage of being frugal with respect to the assumptions they have to make.

References

Atkinson, J. W., & Birch, D. (1970). *The dynamics of action*. New York: John Wiley.

Barber, H. A., & Kutas, M. (2007). Interplay between computational models and cognitive electrophysiology in visual word recognition. *Brain Research Reviews, 53,* 98–123.

Blanton, H., & Jaccard, J. (2006). Arbitrary Metrics in Psychology. *American Psychologist, 61,* 27–41.

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology, 81*, 760–773.

Cohen, J. D., Dunbar, K. & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop Effect. *Psychological Review*, *97*, 332 – 361.

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,* 469–487.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for White Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology*, *36*, 316–328.

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology, 37,* 443–451.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion, 15,* 115–141.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem and stereotypes. *Psychological Review, 102,* 4–27.

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

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,* 17–41.

Herd, S. A., Banich M. T., & O'Reilly R. C. (2006). Neural mechanisms of cognitive control: An integrative model of Stroop task performance and fMRI data. *Journal of Cognitive Neuroscience, 18*, 22–32.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology, 81,* 774–788.

Klauer, K. C., & Mierke, J. (2005), Task-set inertia, attitude accessibility, and compatibility-order effects: New evidence for a task-set switching account of the Implicit Association Test effect. *Personality and Social Psychology Bulletin, 31*, 208–217.

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.

Mierke, J., & Klauer, K. C. (2001). Implicit association measurement with the IAT: Evidence for effects of executive control processes. *Zeitschrift für experimentelle Psychologie, 48*, 107–122.

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

Ortony, A. & Radin, D. I. (1989). SAPIENS: Spreading Activation Processor for Information Encoded in Network Structures. In N. Sharkey (Ed.), *Models of cognition: A review of cognitive science.* Norwood, NJ: Ablex.

Ottaway, S. A., Hayden, D. C., & Oakes, M. A. (2001). Implicit attitudes and racism: Effects of word familiarity and frequency on the implicit association test. *Social Cognition, 19,* 97–144.

Page, M. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23*, 443–512.

Praamstra, P., & Seiss, E. (2005). The neurophysiology of response competition: Motor cortex activation and inhibition following subliminal response priming. *Journal of Cognitive Neuroscience, 17,* 483–493.

Revelle, W. (1986). Motivation and efficiency of cognitive performance. In D. R. Brown & J. Veroff (Eds.), *Frontiers of Motivational Psychology: Essays in honor of J. W. Atkinson*. Berlin: Springer.

Rothermund, K., & Wentura, D. (2001). Figure-ground-asymmetries in the Implicit Association Test (IAT). *Zeitschrift für experimentelle Psychologie, 48*, 94–106.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology, 133*, 139–165.

Schneider, W. & R. M. Shiffrin. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*, 1–66.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127–190.

Snyder-Tapp, K., & Dale, R. (2009). Dynamic competition and the cognitive bottleneck. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.