Modeling the Effect of Evaluative Conditioning on Implicit Attitude Acquisition and Performance on the Implicit Association Test

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

Abstract

Using a previously proposed computational model of human performance on the Implicit Associations Test (IAT), we explore how evaluative conditioning could inform attitude acquisition and formation of automatic associations in memory, and demonstrate the effects of such learning on implicit task performance on the test. This is achieved by augmenting the model with a learning mechanism based on a modified Hebbian learning rule that adapts associative strengths between concepts depending on the temporal proximity of their activation. By manipulating the frequencies at which different stimuli are paired and presented as input to the network, we demonstrate how virtual subjects could acquire associative strengths that were subsequently reflected in simulated IATs as stronger relative preferences in favor of targets that were more frequently presented with positively-valenced stimuli. The model predicts that associations that are already strong have limited prospects for continued reinforcement.

Keywords: Hebbian learning; implicit attitudes; simulation; localist-connectionist networks.

Introduction

Much discussion over the emergence of automatic associations between concepts and their evaluations in memory has taken place within the context of evaluative and classical conditioning (e.g., De Houwer, 2007; De Houwer, Baeyens & Field, 2005; Olson & Fazio, 2001; 2002). Evaluative conditioning is defined as a change in the extent of liking or disliking towards a stimulus that is caused by the frequent pairing of that stimulus with other liked or disliked stimuli (De Houwer, Baeyens & Field, 2005).

The interest in evaluative conditioning research is fueled by the fact that it has the potential to explain the emergence of attitudes and account for the ways in which people's attitudes and beliefs, and consequently their behavior, could be influenced. Thus, it has wide implications especially with regards to consumers' preferences, tastes, and purchasing habits. For instance, Gibson (2008) recently demonstrated the effect of evaluative conditioning in influencing implicit attitudes towards mature brands (e.g., Coke and Pepsi). It was shown that the consistent pairing of positive stimuli with a particular brand could help create and strengthen positive attitudes towards that brand, although the effect was observed only for subjects who had relatively neutral attitudes towards both brands to begin with. Olson and Fazio (2001; 2002) reported similar conditioning effects in which frequent pairings between novel conditioned stimuli (CS) and valenced unconditioned stimuli (US) could result in the acquisition of implicit attitudes towards novel target concepts that were created *a propos* for the experiments, and consequently influence subjects' behaviors and responses on Implicit Association Tests (IAT; Greenwald, McGhee & Schwartz, 1998) involving those novel targets, even though subjects reported no explicit memory of the CS-US pairings.

However, the causal mechanisms by which the evaluative conditioning effect could emerge have yet to be satisfactorily uncovered, owing in part to conflicting empirical data about the conditions under which such effects might occur (De Houwer, Baeyens & Field, 2005). Many controversies revolve around whether associations were learnt as a result of automatic as opposed to conscious controlled processes, whether evaluative conditioning effects were due to a repertoire of processes (as opposed to a single mechanism) or contingent on subjects' awareness of stimuli pairing, and whether the learning is resistant to extinction (De Houwer, 2007; Walther, Weil & Düsing, 2011).

This paper represents our attempts at providing a computational account of the effect of evaluative conditioning on the acquisition of automatic associations between concepts in memory. Through simulations, we examine the impact that frequent pairing of target stimuli with various positively or negatively valenced stimuli would have on implicit task performance, such as on the Implicit Association Test. This is done with a number of goals in mind. First, to provide additional support for the cognitive plausibility of a previously proposed computational model of implicit task performance on the IAT (Quek & Ortony, 2011). Our approach is to augment the localist-connectionist model with a cohesive explanatory account of how automatic associations between concepts in memory could be formed or acquired through experience, a process analogous to how various attitudes are acquired throughout an individual's lifetime.

A second goal is to determine if we could make use of the computational model to address some of the research gaps identified by De Houwer, Baeyens and Field (2005), especially in view of what they see as a lack in the availability of detailed accounts for the processes and mechanisms that underlie evaluative conditioning, and the conditions under which it could occur. More generally, and as pointed out by Van Overwalle and Sieber, (2005), there appears to be limited theoretical advancement in the "understanding of the storage or strengthening of attitude-object associations in human memory." Before more empirical insights are made available, computational approaches such as modeling and simulation could provide an interim but effective means for understanding various candidate processes underlying attitude acquisition or formation (e.g., Eiser, Fazio, Stafford &

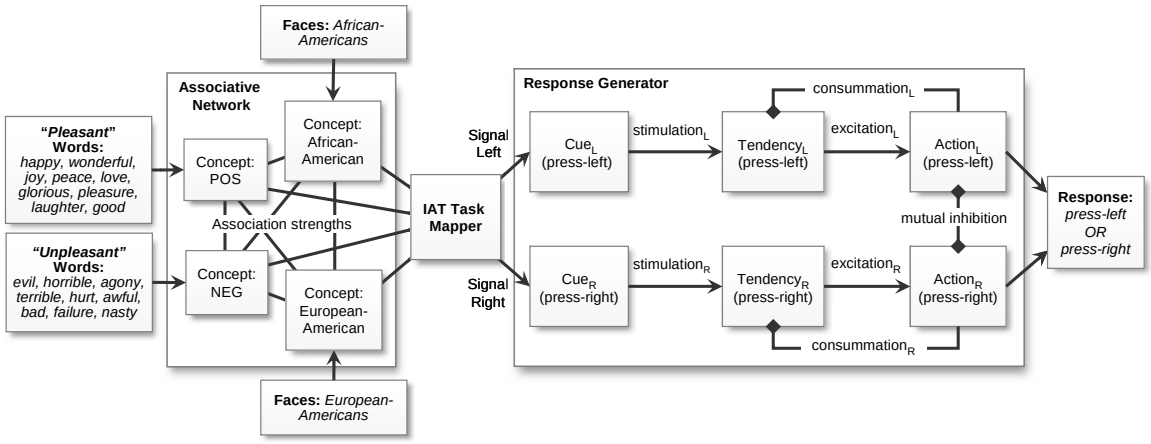

Figure 1. Network model for simulating performance on the IAT (Quek & Ortony, 2011)

Prescott, 2003; Van Overwalle & Sieber, 2005). In our case, a computational model that is demonstrably capable of replicating IAT effects on the basis of different associative strengths between concepts could serve as a platform on which various candidate learning mechanisms could be evaluated, by examining their impact on performance on the IAT. Doing so would also provide an example to demonstrate how learning mechanisms could be incorporated into localist-connectionist models, to fulfill a gap pointed out by some theorists that current associative models of attitudes lack mechanisms that could learn or update internal states and representations in response to information obtained externally from the world, as compared to connectionist models (Van Overwalle & Sieber, 2005). Finally, providing a psychologically plausible mechanism for how associative strengths in the network could be learnt would help allay potential criticisms and modeling concerns over the seemingly arbitrary manner in which associative weights in the earlier model were configured or initialized.

Model Overview

In this section, we provide a brief overview of the computational model used (for more details, see Quek & Ortony, 2011). The model is a localist-connectionist network (e.g., Page, 2000) that emulates the multiple processing pathways from visual perception (i.e., a word or image) to the automatic activation of associated concepts in memory and motor responses. In general, nodes in the network represent concepts while connections represent associations between them. Information is processed in the network through the flow of activation from one node to another, a process governed by the following propagation rule:

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

where x_i is the activation level of a node v_i, $w_{j,i}$ is the weight of the connection $\varepsilon_{j,i}$ from a node v_j which is a neighbor of v_i, E is the set of all edges, α is the propagation gain (set to 0.2) and δ is a decay parameter (set to 0.001) that reduces activation over time. In each time step k, activation spreads to v_i from each of its neighbors v_j at a rate proportional to the weight $w_{j,i}$ of the connection between them. Positive values of $w_{j,i}$ are excitatory while negative values are inhibitory, while a value of zero implies a neutral or null connection.

Model Components

The network comprises a few components (see Figure 1). The *Associative Network* contains nodes representing the target concepts AFRICAN-AMERICAN (AA) and EUROPEAN-AMERICAN (EA), attribute concepts for positivity (POS) and negativity (NEG), input stimuli such as a list of *pleasant* and *unpleasant* words (e.g., *happy, wonderful, joy, evil, horrible, hurt*), and pictures of *European-American* and *African-American* individuals. Connections between target-attribute concept node pairs (i.e., EA↔POS, EA↔NEG, AA↔POS, and AA↔NEG) are taken to represent implicit attitudes. For example, a positive attitude towards EA can be represented as excitatory EA↔POS or inhibitory EA↔NEG associations, or both, such that when EA is activated, POS will be similarly activated while NEG would be inhibited. Similarly, negative attitudes towards EA can be represented by excitatory EA↔NEG or inhibitory EA↔POS associations, or both, such that activation of EA would excite NEG but inhibit POS.

The *Task Mapper* is responsible for transmitting activation from target and attribute concepts to cue_L and cue_R which are nodes indicating that a left or right key-press is required. If the present task requires a right response for "*European-American* or *pleasant*", both POS and EA would be routed to cue_R. These connections are reconfigured at the beginning of each task block, and during which they remain active (see Quek & Ortony, 2011, Figure 2).

The *Response Generator* implements Revelle's (1986) cue-tendency-action model (CTA), which in turn is based on Atkinson and Birch's (1970) dynamics of action theory. CTA describes the dynamic interactions between conflicting tendencies and competing actions. Using CTA as a template, we construct two response-generating pathways (for the left and right key-presses). When activated, response cue nodes will stimulate action-tendency nodes, which will activate response nodes representing the left and right motor responses. When either of the response nodes exceeds a certain activation threshold, it is taken as the winner.

The interactions between the above representations take place as excitations and inhibitions along different propagation pathways. For example, in a task block requiring a left key for "*African-American* or *unpleasant"* and a right key for "*European-American* or *pleasant*", an African-American picture would activate AA, and activation will be transmitted to cue_L. However, if the network is configured with a strong AA↔POS connection, activation will also be transmitted to cue_R, competing with cue_L. This reduces the rate that activation will accumulate in the left response node, and thus a longer time is required for it to reach the response threshold.

Simulating the Implicit Association Test

Each virtual subject's network is first initialized with a set of associative strengths that represents its implicit attitudes, and put through the standard IAT task blocks. The network is provided with a simulated verbal or pictorial input in each trial. The number of iterations taken to produce a response is recorded, and then transformed by a scaling factor into a simulated response time (in milliseconds) of the same order of magnitude as those observed in human subjects (e.g., Greenwald et al., 1998; Klauer, Voss, Schmitz & Teige-Mocigemba, 2007). To compute the IAT effect, we take the raw difference between the simulated mean response times in the two combined task blocks.

Simulating Evaluative Conditioning

To examine the effect that learning processes might have on IAT performance, it would be necessary to extend the localist-connectionist model with mechanisms that could modify its internal features in response to environmental input. While the use of learning is a mainstay of connectionist and parallel distributed processing models (e.g., Cohen, Dunbar & McClelland, 1990; McClelland & Rumelhart, 1986; Read et al., 2010), it is relatively uncommon in localist-connectionist models (Page, 2000).

A number of connectionist models for simulating the automatic acquisition of associations in memory have been proposed (e.g., Eiser, Fazio, Stafford & Prescott, 2003); these typically employ some form of error-correction learning (such as the ubiquitous *delta rule*) that adjusts weights to learn particular stimulus-to-response mappings such that the actual and expected outcomes will eventually converge over time. It is unclear if this is a realistic portrayal of the manner in which associations between concepts are learnt or formed, since the notion of what an *expected outcome* or *reward* ought to be, is ill-defined, or at best, arbitrary. For instance, frequent exposure to a pair of conditioned and unconditioned stimuli need not necessarily involve a motor response or behavioral outcome, though it can be accompanied by a change in state—which in this case would be an increase or decrease in the associative strength between concepts in memory, which can be taken as a change in the degree of liking or disliking for the said stimuli. Work by Herz, Sulzer, Kühn and van Hemmen (1989), and more recently Verguts and Notebaert (2008) employed Hebbian learning rules to learn such state changes.

Hebbian learning (or *plasticity*, Hebb, 1949) can be construed as a form of reinforcement learning in which connections between nodes that *fire* (in the context of neural networks) or are jointly activated within a temporally proximate timeframe would be strengthened over time, such that future joint activation of the associated nodes would co-occur with greater ease. Mathematically speaking, the Hebbian learning rule can be characterized as:

$$\Delta w_{i,j} = \lambda \cdot (x_i \cdot x_j) \tag{2}$$

where λ is a learning rate parameter, x_i and x_j are the activation levels of two nodes v_i and v_j, while $w_{i,j}$ is the weight of the edge $\varepsilon_{i,j}$ originating from node v_i and terminating at v_j. In neural networks, x_i and x_j are known as the pre- and post-synaptic activation levels of the connection between v_i and v_j, respectively. The product $x_i\ x_j$ can be conceived as a measure of similarity between the activation levels of both nodes. The learning rule causes the connection weight between these two nodes to increase proportionately with respect to the degree in which both nodes are temporally activated together. However, this rule is known to be unstable in that connection weights will tend to increase without bounds over time if repeatedly reinforced, or saturate at their maximum and minimum boundaries. To enhance stability, we add a discounting term representing the portion of activation in v_j that is not due to v_i:

$$\Delta w_{i,j} = \lambda \cdot (x_i \cdot x_j) \cdot (x_j - x_i w_{i,j}) . \tag{3}$$

Doing so ensures that $w_{i,j}$ will be adapted in relation to only that portion of the activation in v_j that is not due to v_i, which prevents $w_{i,j}$ from over-learning the joint activation between v_i and v_j. Thus, associations that are already strong to begin with will cease to increase without bounds. Our formulation of the Hebbian learning rule is similar to the simple but provably stable form proposed by Oja (1982):

$$\Delta w_{i,j} = \lambda \cdot x_j \cdot (x_i - x_j w_{i,j}) . \tag{4}$$

The difference between the two formulations is that we have swapped the roles of x_i and x_j within the parentheses, and kept x_i in the product to preserve the role of the similarity term $x_i\ x_j$. Furthermore, we inserted a decay term to allow weights to gradually decay over time, in the absence of activation, to arrive at the following:

$$\Delta w_{i,j} = -\gamma \cdot w_{i,j} + \lambda \cdot (x_i \cdot x_j) \cdot (x_j - x_i w_{i,j}) , \tag{5}$$

where γ is the weight decay rate. For implementation purposes, the learning rule is rewritten as an update function:

$$\begin{aligned} w_{i,j}(k+1) &= (1-\gamma) \cdot w_{i,j}(k) \\ &\quad + \lambda \cdot x_i(k) \cdot x_j(k) \cdot [x_j(k) - x_i(k) w_{i,j}(k)] \end{aligned} \tag{6}$$

We further constrained the model to learn only the weights of associations between positively-activated concept nodes, while allowing associative weights between non-activated or negatively-activated (i.e., inhibited) node pairs to decay and eventually become extinct over time.

Prior to performing the simulation, λ and γ were set to 0.05 and 0.0005 respectively after an initial process of iterative search through parameter space to yield post-learning weights that had a large but unsaturated range.

Simulations

To perform the simulations, we begin with a network configuration in which the weights of the associations EA↔POS, EA↔NEG, AA↔POS, and AA↔NEG are all initialized to zero. In each epoch, 100 pairs of input stimuli, each comprising an attribute concept exemplar (e.g., the word *wonderful*) and a target concept exemplar (e.g., a picture of a White or Black individual) were selected at random. Input nodes corresponding to both exemplars in each stimulus pair were set with an activation of 1.0. The learning rule in Equation (6) was then applied in tandem with the propagation rule defined in Equation (1). Propagation of activation through the network would activate the concept nodes corresponding to these input stimuli. At the same time, the learning rule is expected to enhance the connection weights between pairs of activated concept nodes, for instance, between EA and POS, or AA and POS, using the above example of the word *wonderful* and a picture of a White or Black individual.

By manipulating the frequency at which input exemplars are selected from each attribute and target concept pair, we can simulate situations in which the exemplars of specific target-attribute concept pairs co-occur more frequently than others. As an example, to produce the condition that EA and *pleasant* exemplars co-occur twice as often as EA and *unpleasant*, the frequency for the latter is set to half of the former's. We expect the learning rule to adapt association weights in a manner that will eventually reflect the patterns of distributions across the frequencies at which each target-attribute concept pair is presented.

In this first simulation, two learning conditions were defined, as shown in Table 1. In the first condition (a), the frequency distribution across the target-attribute concept pairs AA+POS, AA+NEG, EA+POS, and EA+NEG were set to 40%, 10%, 10%, and 40%, respectively. The second condition (b) was defined by the distribution 10%, 40%, 40%, and 10%, for target-attribute pairs in the same order. These represent the probability in which paired-stimuli are sampled from the respective concept pair, thus the absolute proportions themselves may vary. Virtual subjects in each condition were put through a pre-learning IAT, followed by the above learning phase during which 100 pairs of stimuli were presented for 100 epochs each. Finally, a post-learning IAT was administered to the virtual subjects. More details concerning the procedures in which the simulated IATs were conducted are found in Quek & Ortony (2011).

Figure 2 shows the evolution of target-attribute associative strengths over the course of learning for 25 virtual subjects in each condition, while the post-learning associative strengths are shown in Table 2. In condition (a), stronger AA↔POS and EA↔NEG associations emerged after learning, while AA↔NEG and EA↔POS increased but at a much slower rate. In (b), stronger associations were found for EA↔POS and AA↔NEG, while the remaining two increased but at a much slower rate. When put through both the pre-learning and post-learning IATs, condition (a) had a non-significant mean IAT effect of -0.03ms prior to learning, $t(24) = -0.132$, $p = 0.896$, but a significant post-learning mean IAT effect of -124.8ms, $t(24) = -19.0$, $p < .001$, which is indicative of an implicit preference for AA over EA. Similarly, condition (b) exhibited a non-significant pre-learning mean IAT effect of 0.24ms, $t(24) = 0.771$, $p = 0.448$, but a significant post-learning mean IAT effect of 120.6ms, $t(24) = 22.3$, $p < .001$, indicative of an implicit preference for EA over AA. Considering that each network began with non-significant pre-learning IAT test scores but expressed significant post-learning IAT effects, and since no other modifications were made to the network, we may conclude that the increase in IAT effect is due to the associations that were acquired over the course of learning. As expected, the emerging associative strengths in each condition (Table 2) showed a similar pattern to the distributions of presentation frequencies of the corresponding target-attribute pairs (Table 1).

To investigate the impact of different co-occurrence frequencies on the post-learning IAT effect, we repeated the above simulation for 250 virtual subjects, only this time varying the frequency distribution for each subject by interpolating randomly between 50%, 0%, 0%, 50%, and 0%, 50%, 50%, 0% for the respective target-attribute concept pairs AA+POS, AA+NEG, EA+POS, and EA+NEG that were presented during learning. When the proportions of both AA+POS and EA+NEG stimuli were reduced from 50% to 0%, the proportions of AA+NEG and EA+POS stimuli were increased from 0% to 50%, in a complementary manner, while ensuring that all four proportions add up to 100%.

Table 1: Presentation frequency of paired stimuli in two experimental conditions during the learning phase

Stimulus Pair	Prototypical exemplars	Presentation Frequency	
		Condition (a)	Condition (b)
AA+POS	 + "*happy*"	40%	10%
AA+NEG	 + "*sorrow*"	10%	40%
EA+POS	 + "*laughter*"	10%	40%
EA+NEG	 + "*horrible*"	40%	10%

Note: EA: European-American; AA: African-American; POS: Positivity; NEG: Negativity.

Table 2: Post-learning target-attribute associative strengths

Association	Condition (a)		Condition (b)	
	M	*SD*	*M*	*SD*
AA↔POS	.865	.083	.287	.148
AA↔NEG	.289	.084	.862	.093
EA↔POS	.256	.120	.862	.091
EA↔NEG	.874	.061	.292	.117

Note: EA: European-American; AA: African-American; POS: Positivity; NEG: Negativity.

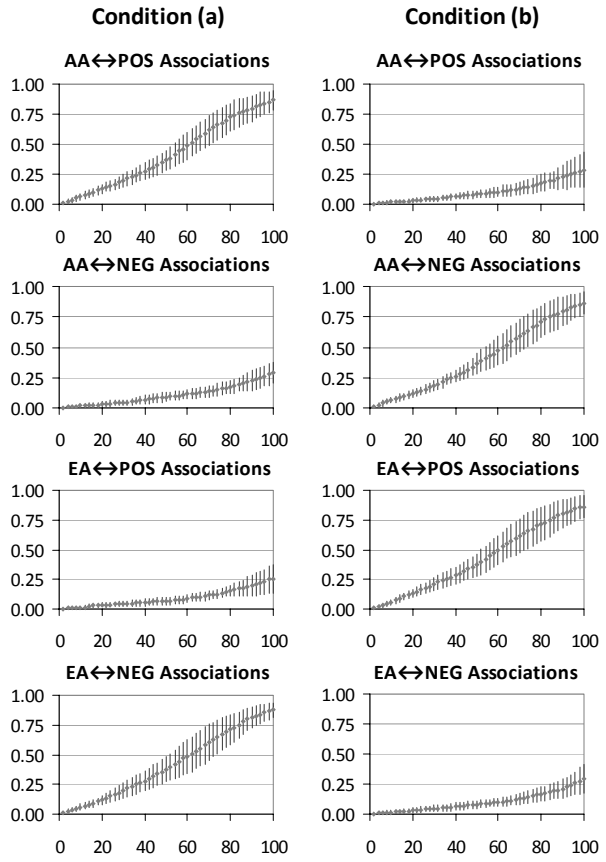

Figure 2. Evolution of associative strengths over the course of learning, for virtual subjects in two conditions. EA: European-American; AA: African-American; POS: Positivity; NEG: Negativity. Y-axis: associative strengths. X-axis: learning epochs. Error bars: standard deviations.

Plotting the post-learning IAT effect against the presentation frequencies for the four target-attribute concept pairs in Figure 3, we found that when a larger proportion of EA+POS and AA+NEG paired stimuli were presented to the model during the learning phase, the post-learning IAT subsequently produced larger IAT effects that were in favor of EA. Conversely, when more input stimulus pairs were selected from AA+POS and EA+NEG and presented to the model during learning, the post-learning IAT had larger IAT effects in favor of AA. When all input stimulus pairs were presented with about the same probability (i.e., keeping the proportions to 25% for each target-attribute concept pair), the post-learning IAT effect was close to zero.

Discussion

With the computational model, we have demonstrated how automatic associations between target and attribute concepts could be acquired by repeated exposure to pairs of input exemplars—as similarly achieved in human subjects via evaluative or classical conditioning (De Houwer, 2007; Olson & Fazio, 2001). Stronger associations were acquired for target-attribute concept pairs whose input exemplars were presented together more frequently, and weaker associations were learnt for other target-attribute concept pairs whose input exemplars were presented together less frequently.

These simulations have some important implications especially with regards to the malleability of implicit attitudes. First, the ability to influence or generate novel associations through consistent pairing of target and attribute stimuli supports the findings of Olson and Fazio (2001) and of Gibson (2008), particularly the latter's discovery that the effects of evaluative conditioning were observed only for subjects who initially had relatively neutral attitudes towards the targets, and not those who already possess a significantly stronger preference for one target over the other. In our terms, this could be explained by the longer amount of time required for stronger associative strengths to decay or weaken over time when the corresponding paired stimuli were no longer presented as frequently.

Second, the evolution of associative strengths over learning epochs in Figure 2 showed a gradual slowdown as they approached 1.0, suggesting that, as these associations increase in strength over the course of learning, the extent to which they can be further increased is limited. Thus, there is limited room for the continued positive reinforcement of associations whose strengths are already high, such that they become less susceptible to learning. Consistent with empirical observations (Gibson 2008; Joy-Gaba & Nosek, 2010), the model thus predicts that this would limit the impact that evaluative conditioning might have on attitudes that have already been firmly ingrained, and thus the continued malleability of attitudes through such means could be reduced. While it could be argued that this effect is largely a result of the modified Hebbian learning rule we devised in Equation (6) that limits the extent to which already-strong associations could continue to be increased, the weights will nonetheless be subject to the finite upper boundary even when the standard unconstrained Hebbian rule in Equation (2) were used instead, and give rise to the same observations.

Third, the simulation results so far are in agreement with Mitchell, Anderson and Lovibond's (2007) proposal that the IAT itself could be used as a means for detecting the occur-

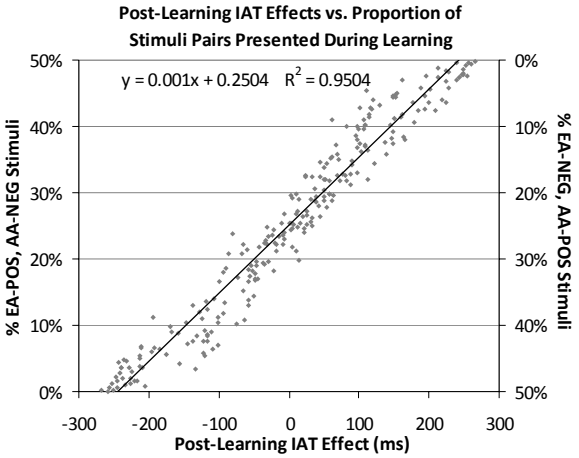

Figure 3. Post-learning IAT effects in virtual subjects (N=250) across presentation frequencies of input stimuli from each of the target-attribute concept pairs during the learning phase.

rence of evaluative conditioning, which, to be consistent with Gibson (2008), is to be expected only for target concepts that have yet to be strongly associated with any particular attributes. Finally, considering that the simulated mechanisms of learning are not specific to valenced attributes, they could be relevant not just for evaluative conditioning, but for explaining other more generic forms of conditioning or learning, such as the effectiveness of reaffirmations to enhance self-concept and self-esteem.

Conclusion

In summary, we have augmented the cognitive plausibility of our computational model (whose purpose was to account for the emergence of IAT effects) by providing a cohesive and cognitively-plausible account of the manner in which implicit attitudes could be acquired through evaluative conditioning, as well as their subsequent effects on implicit task performance on a simulated IAT. This is achieved via a modified Hebbian learning rule that adapts associations between concept representations in memory relative to the different frequencies at which target stimuli are paired with other positively or negatively valenced stimuli. An additional contribution of the model is in demonstrating how localist connectionist models too are amenable to learning mechanisms, just like their connectionist counterparts (Van Overwalle & Sieber, 2005). Extending the simulations beyond the permitted scope of this paper to include additional learning conditions and a more comprehensive analysis of the viability of the learning algorithms presented (in comparison to possibly other candidates) would be a logical continuation of this work in future.

Acknowledgments

We wish to thank the anonymous reviewers for their valuable feedback and suggestions. B.-K. Quek is supported by a postdoctoral fellowship from the Agency for Science, Technology and Research (A*STAR), Singapore.

References

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

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.

De Houwer, J. (2007). A Conceptual and Theoretical Analysis of Evaluative Conditioning. *The Spanish Journal of Psychology*, *10*, 230–241.

De Houwer, J., Baeyens, F., & Field, A. P. (2005). Associative learning of likes and dislikes: Some current controversies and possible ways forward. *Cognition and Emotion*, *19*, 161–174.

Eiser, J. R., Fazio, R. H., Stafford, T., & Prescott, T. J. (2003). Connectionist simulation of attitude learning: Asymmetries in the acquisition of positive and negative evaluations. *Personality and Social Psychology Bulletin, 29,* 1221–1235.

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research*, *35*, 178–188.

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.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley & Sons.

Herz, A., Sulzer, B., Kühn, R., & van Hemmen, J. L. (1989). Hebbian learning reconsidered: Representation of static and dynamic objects in associative neural nets. *Biological Cybernetics*, *60*, 457–467.

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology*, *41*, 137–146.

McClelland, J. L., & Rumelhart, D. E. (1986). A distributed model of human learning and memory. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. II. Psychological and Biological Models (pp. 170–215)*. Cambridge, MA: MIT Press/Bradford Books.

Mitchell, C. J., Anderson, N. E., & Lovibond, P. F. (2003). Measuring evaluative conditioning using the Implicit Association Test. *Learning and Motivation*, *34*, 203–217.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*, 267–273.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological* Science, *12*, 413–417.

Olson, M. A., & Fazio, R. H. (2002). Implicit acquisition and manifestation of classically conditioned attitudes. *Social Cognition*, *20*, 89–103.

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

Quek, B.-K., & Ortony, A. (2011). Modeling underlying mechanisms of the Implicit Association Test. In L. Carlson, C. Hoelscher, & T.F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp.1330–1335). Austin, TX: Cognitive Science Society.

Read, S. J., Monroe, B. M., Brownstein, A. L., Yang, Y., Chopa, G., & Miller, L. C. (2010). A neural network model of the structure and dynamics of human personality. *Psychological Review, 117*, 61–92.

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.

Van Overwalle, F., Siebler, F. (2005). A connectionist model of attitude formation and change. *Personality and Social Psychology Review*, *9*, 231–274.

Verguts, T., & Notebaert, W. (2008). Hebbian learning of cognitive control: Dealing with specific and nonspecific adaptation. *Psychological Review*, *115*, 518–525.

Walther, E., Weil, R., & Düsing, J. (2011). The role of evaluative conditioning in attitude formation. *Current Directions in Psychological Science*, *20*, 192–196.