


# Modeling the Dynamics of Evaluation: A Multilevel Neural Network Implementation of the Iterative Reprocessing Model

Personality and Social Psychology Review  
2015, Vol. 19(2) 148–176  
© 2014 by the Society for Personality  
and Social Psychology, Inc.  
Reprints and permissions:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/1088868314544221  
pspr.sagepub.com  


Phillip J. Ehret<sup>1</sup>, Brian M. Monroe<sup>2</sup>, and Stephen J. Read<sup>3</sup>

## Abstract

We present a neural network implementation of central components of the iterative reprocessing (IR) model. The IR model argues that the evaluation of social stimuli (attitudes, stereotypes) is the result of the IR of stimuli in a hierarchy of neural systems: The evaluation of social stimuli develops and changes over processing. The network has a multilevel, bidirectional feedback evaluation system that integrates initial perceptual processing and later developing semantic processing. The network processes stimuli (e.g., an individual's appearance) over repeated iterations, with increasingly higher levels of semantic processing over time. As a result, the network's evaluations of stimuli evolve. We discuss the implications of the network for a number of different issues involved in attitudes and social evaluation. The success of the network supports the IR model framework and provides new insights into attitude theory.

## Keywords

attitudes, automatic/implicit processes, person perception, social cognition, social neuroscience, neural networks, computational modeling, evaluation

When we encounter an individual, how do we form an attitude toward that person? Social psychologists have been fascinated with dual process and dual system theories of attitudes as potential answers to that question, and indeed, these theories have enabled significant research advances in the fields of stereotyping, person perception, and attitudes (for a review, see Gawronski & Creighton, 2013). Although there are many important distinctions between dual process theories and dual system theories, as well as within each theoretical perspective, generally speaking, all of these theories assume that attitudes are represented and/or processed by *two distinct* systems or collections of processes: One that is quick and (relatively) uncontrolled and a second that is slower and involves controlled and deliberative processing. Much of the interest in the quick and uncontrolled route is with the possibility that these relatively automatic attitudes reveal how an individual “really feels” without the intervention of self-presentation or self-regulatory processes.

Although dual process and dual system theories have contributed considerably to our understanding of attitudes and attitude formation, recent theorizing and empirical evidence suggest that the dual process and dual system models, particularly in their strong form (i.e., dual *and largely independent* processes or systems), may need to be revised. There are good reasons to think that attitudes toward a target develop dynamically over time as the result of ongoing interactions among

multiple brain systems. For example, Cunningham and his colleagues (Cunningham & Zelazo, 2007; Cunningham, Zelazo, Packer, & Van Bavel, 2007), in their iterative reprocessing (IR) model, have argued that attitudes and evaluations develop as the result of an iterative process of interaction between neural systems in which evaluations develop over time as activation spreads both forward and backward across a hierarchy of neural systems. The goal of the current article is to provide an implemented neural network model of this process that shows that evaluation over time can be captured with a dynamic, interaction process that does not require two independent processes or systems.

First, we broadly discuss dual process and dual system theories, and then review theoretical criticisms of them. Second, we review evidence that strong dual process or dual system theories have difficulty explaining. Third and finally, we argue for the utility of the IR model and provide support

<sup>1</sup>University of California, Santa Barbara, USA

<sup>2</sup>University of Alabama, Tuscaloosa, USA

<sup>3</sup>University of Southern California, Los Angeles, USA

## Corresponding Author:

Stephen J. Read, Department of Psychology, University of Southern California, 3620 McClintock Ave, Los Angeles, CA 90089-1061, USA.  
Email: read@rcf.usc.edu

for this theory with a neural network that provides a concrete implementation of this alternative theoretical framework.

## Dual Process and Dual System Theories

The most prevalent accounts of attitudes as the result of two distinct processes or systems have been provided by various dual process and dual system models (e.g., Fazio, 1995; Gawronski & Bodenhausen, 2006; Greenwald & Banaji, 1995; Strack & Deutsch, 2004; for overviews, see Evans, 2003; Kruglanski & Orehek, 2007). These models posit that the first process or system is quick and automatic, resulting in implicit attitudes. The second process or system is relatively slower and more deliberative, resulting in explicit attitudes. These two processes or systems have been variously named (e.g., impulsive vs. reflective, System 1 vs. System 2, automatic vs. controlled). Although the precise nature of each process or system differs from account to account, all share the basic two features of an initial quick, automatic process/system and a later more controlled and deliberative process/system. One attraction of these dual process and dual system models is that they explain how individuals can have implicit and explicit attitudes about the same target that are very different from one another.

In addition to distinguishing between automatic and controlled processes, several researchers (e.g., Gawronski & Bodenhausen, 2006; Strack & Deutsch, 2004) have claimed that automatic or impulsive processes are associative (based on similarity and contiguity, and independent of truth value), whereas reflective or controlled processes are based on propositional processes (which have truth values). They have used this distinction between associative and propositional processes to explain the conditions under which implicit and explicit attitudes may differ.

## Criticisms of Dual Process and Dual System Theories

A number of theorists (e.g., Evans, 2008; Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011; Kruglanski & Orehek, 2007), although not denying the empirical reality of the distinction between such early and later developing evaluations, have argued against the basic idea of dual process and dual system models and their explanatory value in understanding the differences between implicit (early) and explicit (later) evaluations. For example, Keren and Schul (2009) have argued that the whole enterprise of dual system models is fundamentally flawed. They note that different dual system models make somewhat different assumptions about the characteristics of the two systems and that the features that theorists argue are critical to defining the two systems do not neatly separate into two groups, but instead overlap and conflict. They suggest that if researchers cannot even agree on what features distinguish the two systems, that perhaps the distinction does not make sense.

Evans (2008) questioned the usefulness of the dual process distinction. Although he agrees that it might make sense to talk about a reflective or System 2 process that is responsible for controlled, deliberative thinking, he notes that rather than there being a single impulsive or System 1 process, that there are actually a number of different systems that handle such things as vision, audition, language, memory, emotion, and motivation.

Kruglanski and Dechesne (2006) also argued against the idea of dual process models, and as an alternative, Kruglanski proposed a unimodel (e.g., Kruglanski & Gigerenzer, 2011). According to Kruglanski, what other researchers argue are two different processes is instead a single rule-based process, where various rules are used to make inferences from evidence. What might look like different processes is simply the use of different rules with different parameters.

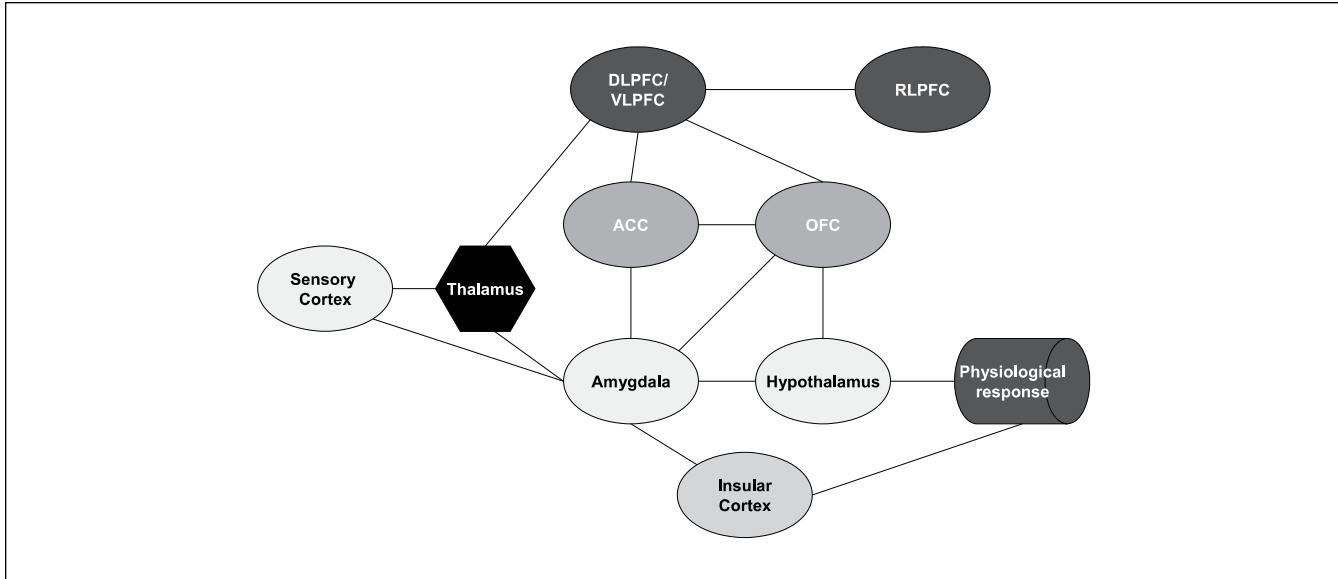
## Evidence of Variability of Implicit Attitudes

Researchers interested in implicit attitudes have also begun to demonstrate that implicit attitudes are not necessarily stable, “true” attitudes in the sense that they are free from situational influences. Although not all dual process or dual system models ignore the impact of situation or context on implicit attitudes (e.g., Fazio, 2007), many dual process or dual system models do not discuss the influence of context on attitudes. Research has demonstrated that implicit/automatic evaluations and attitudes can be influenced by a number of situational factors such as context (for reviews, see Blair, 2002; Gawronski & Sritharan, 2010). One example of research showing these effects demonstrated that context can moderate automatic and controlled racial bias when a picture of a context was continuously presented on the screen during the critical phase of an evaluative priming procedure (Barden, Maddux, Petty, & Brewer, 2004). Situational context has also been found to influence automatic group attitudes and stereotypes (Wittenbrink, Judd, & Park, 2001).

Important to recognize, and a point to which we will return later, is that in each of these experimental paradigms, the moderator being investigated, such as context, is presented *before* the individual being evaluated. Thus, in these examples of the impact of context, contextual information precedes individual level information, and results in changes to automatic or implicit attitudes. This research calls into question some dual process and dual system models’ argument that automatic or implicit attitudes are just recalled or activated attitudes. Instead, we suggest that a more dynamical understanding may better account for the moderating effects of other factors such as context on evaluation.

## Evidence for Dynamical Processing

Partially in response to the criticisms of dual process and dual system models, as well as to new experimental techniques



**Figure 1.** Figure from Cunningham, Zelazo, Packer, and Van Bavel (2007). This presents a simplified neural circuit for the IR model. Lines in the diagram represent the major proposed connections, with most being bidirectional. Processing starts at the sensory cortex. OFC = orbito-frontal cortex; ACC = anterior cingulate cortex; VLPFC = ventrolateral prefrontal cortex; DLPFC = dorsolateral prefrontal cortex; RLPFC = rostralateral prefrontal cortex.

(e.g., functional magnetic resonance imaging [fMRI], mouse tracking software), research has emerged suggesting that social cognition is highly dynamic and integrative (e.g., Freeman & Ambady, 2011; Wojnowicz, Ferguson, Dale, & Spivey, 2009). Recent research has used fMRI and electroencephalography (EEG) to show how self-categorization and social identity can influence person perception and evaluation dynamically over time (for a review, see Van Bavel, Xiao, & Hackel, 2013). Other work by Freeman and Ambady (2011) utilized a neural network and mouse tracking software to support the validity of a dynamical approach to person categorization. Research investigating own-race bias has used EEG to show that differential response patterns based on motivational states and target race emerge only 100ms after stimulus onset, well before any executive control can influence this “automatic” process (Cunningham, Van Bavel, Arbuckle, Packer, & Waggoner, 2012). This dynamic perspective stands in contrast to dual process and dual system models that generally do not provide a mechanism to explain the variability in automatic attitudes, and it particularly contrasts with the strong models that do not posit much, if any, interaction between the dual processes or systems (for additional discussion, see Van Bavel, Xiao, & Cunningham, 2012; Van Bavel et al., 2013). However, this does not mean that all dual process and dual system models fail to account for time in the formation of attitudes (for models where attitudes are formed through a series of stages or sequential processes, see Brewer & Feinstein, 1999; Fiske, Lin, & Neuberg, 1999), or claim that there is no dynamic interaction between different processes (e.g., Fiske et al., 1999).

## The IR Model

Cunningham’s (Cunningham & Zelazo, 2007; Cunningham et al., 2007) proposed IR model argues that evaluation is the result of an iterative reprocessing of information that develops over time across multiple interacting brain systems. Thus, rather than there being a distinct number (e.g., two) of categorical processes, there are multiple brain systems that interact over time to form attitudes. For example, in developing an evaluative response to a Black male doctor, there may be an earlier evaluation based on early perceptual analyses of race and gender (e.g., Ito & Urland, 2003), followed by more detailed processing involving higher-level semantic information: The information about his clothing and the situational context will lead to processing of the concept of doctor, and then the concept of doctor may lead to processing of associated attributes such as intelligent, caring, and popular, which in turn lead to associated evaluations. As a result, the evaluation of the individual will develop over time as different brain systems are brought online and continually interact and mutually influence each other to process and integrate different information (see Figure 1).

Importantly, we are not suggesting that earlier and later processing necessarily represent automatic and controlled processing, respectively. Rather, we argue, in accordance with the IR model, that evaluations are continuously evolving over time and result from the interactions of multiple systems. Given this perspective, differences in outcomes between earlier and later processing are not necessarily due to automatic versus controlled processes. Recent research

with the IR model shows that context and motivation can influence extremely early (around 100 ms after stimulus presentation) evaluations (Cunningham et al., 2012). This is an important finding, as it helps to explain the body of research indicating that factors such as context can moderate “automatic” attitudes (e.g., Barden et al., 2004). Moreover, the fact that some studies find changes in evaluation using putatively automatic measures is not really informative about the time course of processing. For example, Implicit Association Tests (IATs) are typically administered well after stimulus processing is finished, and implicit measures such as the IAT or the evaluative priming task may not even be able to capture these earlier evaluations, as around 300 ms or higher is typically set as a lower bound for reaction times (Greenwald, McGhee, & Schwartz, 1998).

In this article, we present a connectionist account of important aspects of the IR model. We focus on the role of earlier perceptual evaluation and the later processing of semantic information as the perceiver arrives at a more detailed representation of the target. We show how earlier evaluations of a target may change over time as further information is processed and integrated with the developing representation, and we do so without postulating different kinds of representations, such as distinct implicit or explicit attitudes, or different kinds of processes, such as associative or propositional processes.

The current neural network captures important phenomena that have been taken as evidence for two distinct attitudinal processes or systems. For example, researchers frequently take the dissociation between implicit and explicit measures of attitudes as evidence for two processes or systems. In contrast, we argue that this “dissociation” is not due to two distinct processes/systems but is the result of dichotomizing the continuous evolution of the interaction of multiple neural systems. Differences between “implicit” and “explicit” attitudes can arise from measuring only two time points of the time course of evaluation as additional information is processed. One of the central points we will demonstrate is how an individual can have an early attitude toward a target that evolves over time to become quite different. For example, we show how one could have an early attitude toward a Black doctor that is quite negative because one has a negative attitude toward Black males, but the attitude changes over time to become quite positive as information about the target is more fully processed. And conversely, we show how someone could have an early attitude toward a White gang member that is quite positive, because of an early positive attitude toward White males, but that the attitude can change with further processing to become quite negative.

## Connectionist Models

By constructing a neural network model that is based on biological principles and that provides a computational implementation of the IR model, we aim to add additional support

for the validity of the IR model in explaining the formation of attitudes. Connectionist models function by the spread of activation over time across weighted links among networks of nodes that represent features and concepts. In these networks, the spread of activation develops over time, and as a result, the development of mental representations and associated evaluation also occurs over time. Initial features and concepts that are activated may have different associated evaluations than the subsequent, more detailed representation.

The development of representations in connectionist models is a function of the pattern of activation across the nodes in the network. Different input activations due to different contexts, as well as differences in activation of nodes due to chronic and temporary accessibility, will lead to different patterns of activation and thus to different representations for a set of inputs. Several researchers (Bassili & Brown, 2005; Conrey & Smith, 2007; Monroe & Read, 2008; Read & Monroe, 2009; Smith & Conrey, 2007) have presented conceptual analyses of the implications of such a connectionist approach for understanding the development of evaluation over time.

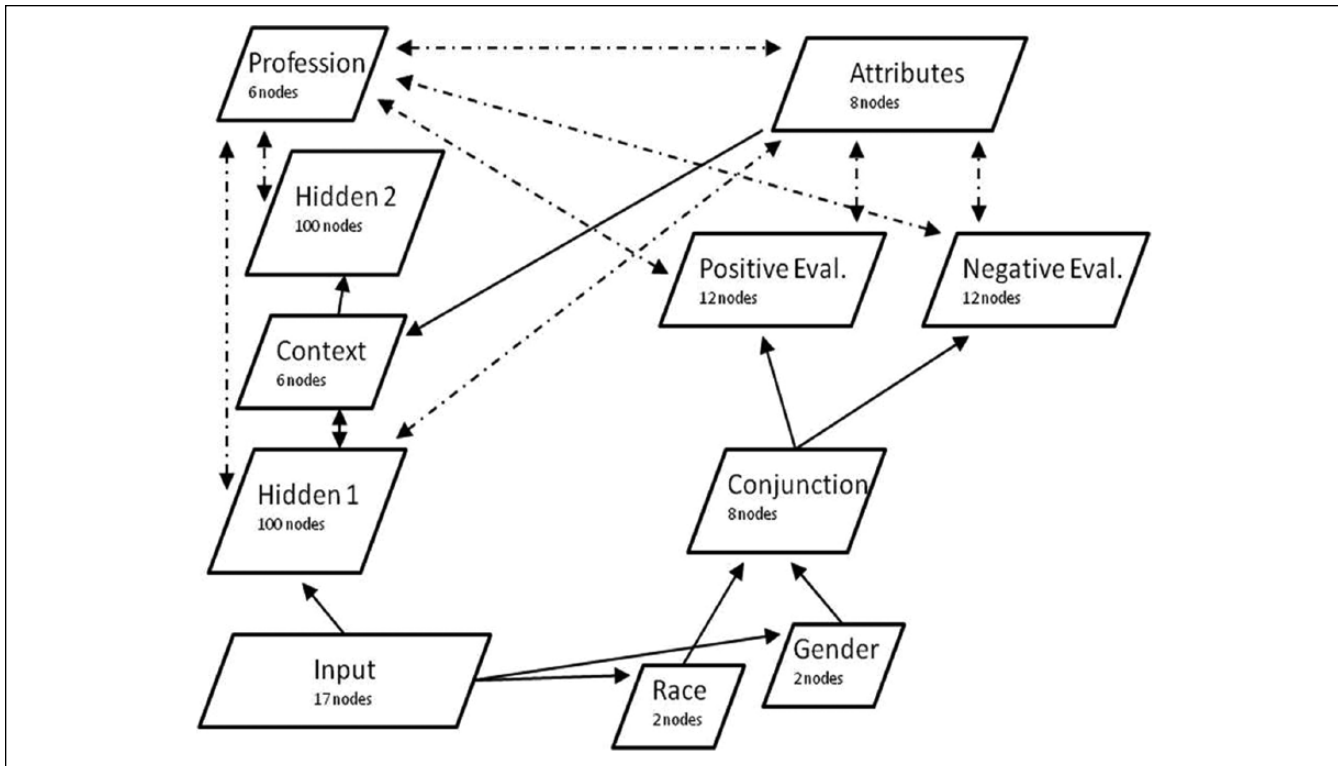
A connectionist approach can also be used to explain the sensitivity of evaluation and attitudes to context that has been shown in several studies (e.g., Correll, Wittenbrink, Park, Judd, & Goyle, 2011; Wittenbrink et al., 2001). For instance, Correll and colleagues (2011) showed that evaluations of a Black male were influenced by the context in which the Black male was seen: A Black male was perceived as less threatening in an upscale neighborhood as opposed to a dangerous location such as a dilapidated building.

Monroe and Read (2008) used their attitudes as constraint satisfaction (ACS) model to simulate some aspects of the time course of attitudes. However, they did not attempt to capture the more detailed neural structure outlined in the IR model. Read and Monroe (2009) presented a conceptual outline of a connectionist model that addresses some aspects of the IR model, particularly the role of multiple layers of processing, with early layers doing largely sensory processing and later layers doing semantic processing. However, they did not present an actual model or simulations.

## The Network

### Overview

Our neural network seeks to implement central aspects of the IR model in a theoretically and biologically plausible framework. The neural network is governed by mathematical functions that seek to represent actual neural activity (e.g., action potentials, bidirectional connections; see O'Reilly & Munakata, 2000, for more details); thus, the network seeks to model, in a simplified manner, actual neural processes. In the specific implementation shown here (see Figure 2), the network is presented with observations of 16 “individuals” consisting of information about features such as skin tone,



**Figure 2.** Neural network layers and connections.

Note. Dashed lines represent bidirectional connections. Solid lines represent unidirectional connections.

gender-specific features, clothing, and aspects of their physical environment. From these observations, the network quickly identifies race, either Black or White, and the gender of the individual, which activates earlier evaluations associated with these two features. The earlier evaluation based on race and gender is part of the first iterations of the stimuli processing and likely occurs unconsciously in limbic structures such as the amygdala (Knutson, Mah, Manly, & Grafman, 2007; Phelps et al., 2000). This is consistent with research on early perceptual cue processing that has demonstrated that non-Black participants preferentially direct attention to Black targets very early in processing (by about 120 ms after stimulus onset) and that attention to gender emerges soon thereafter (about 180 ms; Ito & Urland, 2003).

As activation continues to spread through the network, the stimuli are reprocessed in higher-order semantic layers that recognize higher-level concepts such as profession and physical location (e.g., office or street). Processing of these higher-level concepts leads to further associated positive and negative evaluations. Once the individual's profession or role is determined, stereotypic attributes such as "intelligent" for a doctor or "violent" for a gang member are activated. These attributes then lead to associated positive or negative evaluations, leading to continued revisions of positive and negative evaluations. In the current network, later processing largely involves semantic information, but we should note

that the IR model does not restrict later processing to just semantic information.

The two separate positive and negative evaluation layers allow for tracking of the separate evolution of these two components of evaluation, eliminating the limitation of a single evaluation continuum that cannot capture ambivalence. Moreover, this assumption is consistent with research demonstrating the existence of two separate evaluation systems (Cacioppo & Berntson, 1994; Cacioppo, Gardner, & Berntson, 1997). Although we have two evaluation layers, this is more an artifact of the constraints of the modeling program, and these two layers may be more accurately thought of as representing two pools of neurons associated with positive and negative evaluations, respectively.

### *Predictions and Manipulations*

The overarching goal is to use the proposed network to examine examples of the continuous development of evaluation from the earlier processing of racial and gender stereotypes, based on analysis of perceptual cues, to the later development of evaluations that integrate occupational and contextual perceptual cues. Specifically, we present the network, which is trained to hold a negative Black stereotype, with both stereotype consistent individuals (e.g., Black gang member) and stereotype inconsistent individuals (e.g., Black doctor). We

selected these individuals in part based on previous research investigating the interaction between stereotypes and other factors such as traits in influencing evaluations (Kunda & Thagard, 1996; Sinclair & Kunda, 1999).

The dependent variable for all the predictions is the time course of evaluation. As a result, the predictions are all quite distinct from predictions that would be made by dual process and dual system models. Dual process and dual system models that argue for distinct and non-interactive processes/systems would generally predict two distinct attitudes, an automatic or implicit attitude and an explicit or controlled attitude (e.g., Wilson, Samuel, & Schooler, 2000). Other dual process or dual system models that posit more interaction between the two processes/systems still generally predict two attitudes, but allow for the later processing to influence the earlier automatic attitude in a corrective fashion (e.g., Wegener & Petty, 1997). Although each specific dual process or dual system model will likely have differences in the specifics of their predictions or mechanisms underlying their predictions, none posits a dynamical evolution of attitudes as argued by the IR model.

Specially, there were four main predictions. To test these predictions, the network first learned evaluations of four types of individuals: Black males, Black females, White males, and White females, and then it learned more detailed information about 16 different types of individuals. This sequence of learning had no theoretical significance, but was done simply to ensure that the network properly learned both the simple gender and race related stereotypes and the more detailed representations.

The first prediction was that evaluations would dramatically shift over the time course of evaluation when earlier stereotype evaluations conflict with semantic information processed by later iterations of the network. To test this prediction, the network initially learned the evaluations for four types of individuals. These individuals were represented only in terms of gender and race cues to represent positive and negative racial stereotypes. Thus, the network learned highly negative evaluations of Black males, somewhat less negative evaluations of Black females, highly positive evaluations of White males, and slightly less positive evaluations of White females. Following this early learning, the network learned associated evaluations for a number of individuals, such as a Black and White male doctor, a Black and White female gang member, and so on. Following this later learning, we then tested the network with a number of different individuals and tracked the evolution of their evaluation over time, from the early cycles of processing that are based largely on perceptual cues of race and gender to the later cycles of processing that are based on information about context and profession, and their associated attributes. We expected that for some of these individuals, there might be fairly dramatic shifts over the time course of evaluative processing. For example, we predicted that earlier evaluations of the Black male doctor would be highly negative because

earlier processing would rely almost entirely on the race and gender cues, whereas later processing would rely increasingly more on information about profession, context, and related attributes as processing continued. Conversely, for White male gang members, we expected earlier perceptual processing to result in a largely positive earlier evaluation, followed by an increasingly negative later evaluation based on continued processing that incorporated evaluations of their profession and observed context.

In addition to these basic simulations, we conducted three follow-up simulations to test the remaining three predictions. For the first follow-up simulation, we manipulated the strength of the initial negative stereotype of Black males and females to test the second prediction that the strength of the basic stereotype (strong vs. moderate vs. low) would influence how the evaluation of the individuals would evolve over time. We predicted that as the initial negative stereotype became weaker, that the degree of earlier negative evaluation would decrease. This simulation provides insight into how varying stereotype strengths can affect the evolution of evaluation of relevant targets.

Assuming that we would see a dramatic shift in the evaluation of the Black doctors from earlier negative evaluation to later positive evaluation, we conducted a second follow-up simulation to test the third prediction that the shift in evaluations from earlier negative evaluation to later positive evaluation would depend on previous experience with Black doctors. Specifically, we were interested in whether you might still get the same kind of shift, albeit weaker, if you never had experience with Black male doctors but still had considerable experience with White doctors. This particular simulation provides insight into how different sources of information are integrated when a novel type of individual is evaluated. We predicted that when presented with a novel individual (in the follow-up simulation, a Black male doctor that was not included in the learning stimuli), the network would successfully integrate the information, resulting in more positive later evaluations. However, we did expect that during the evaluation time course, the strength of the positive evaluation would always be less positive than for the case where the network had experience with a Black male doctor, as the network would more heavily rely on racial stimuli and the associated negative evaluations.

In the third follow-up simulation, we tested the fourth prediction that if context cues were encountered before the race and gender cues, this context information would modify the evaluative responses to the race and gender cues.<sup>1</sup> So in the case of the Black male doctor, for example, if the network processed context cues (i.e., physical location and clothing) immediately prior to processing race and gender cues, the network would have limited or no negative evaluation over the evaluation time course. The purpose of this simulation is to demonstrate that the network can indeed integrate information activated prior to the processing of race and gender cues, consistent with research that has demonstrated that

very early processing can be influenced by factors such as context (e.g., Barden et al., 2004; Cunningham et al., 2012).

### Architecture

The neural network (see Figure 2) was constructed with *emergent 5.1.0* software using the Leabra (local, error-driven and associative, biologically realistic algorithm) architecture (for a detailed description, see Aisa, Mingus, & O'Reilly, 2008; O'Reilly & Munakata, 2000). The Leabra architecture allowed us to model several important processes central to the IR model. First, it allowed for the computation and tracking of evaluations over time as a result of the processing in layers and the interaction of multiple network layers. Second, it allowed for separate positive and negative evaluation layers. Although Leabra's architecture is biologically based, we do not claim that the current network is an exact analog of specific neural pathways. Instead, the network is intended to demonstrate central theoretical characteristics of the IR model in a simplified, yet biologically plausible, computational neural network.

Processing in a neural network occurs by the flow of activation among the nodes in the network in response to the presentation of input patterns to the network. The *network architecture* determines how many nodes there are, how they are arranged, and how they are connected. The *activation function* determines how activation propagates through the network by taking the activation of all the nodes coming into a node, integrating those activations, and then computing the output activation of the node as a function of the inputs. The pattern of connections among nodes can change with experience and learning. The *learning rule* dictates how connections are changed during learning, typically as a function of the patterns of associations among the activations of nodes. Specific network details and more details about the Leabra architecture (i.e., inhibition and activation settings, learning settings) are provided in Appendix A.

**Current network.** The network has 11 layers, with layers for stimulus inputs, race and gender recognition, race and gender conjunctions, input conjunction (hidden layer), four higher-order semantic layers (context, profession, attributes, and one hidden layer), and positive and negative evaluation layers. Hidden layers learn to represent conjunctions or combinations of features and do not get input from the environment nor do they directly get a teaching signal from the environment. Because their function is to re-represent information so that another system can utilize it later in the processing stream, they are often operating "behind the scenes" to enable other processes of interest; thus, they are commonly referred to as "hidden" layers. Connection directions between all layers of the network vary (see Figure 2).

**Layers.** The input layer is a localist layer consisting of 17 nodes, each representing a visual stimulus feature (e.g., skin color, clothing; see Table 1). Both the race and gender layer

**Table 1.** Layers' Node Representations.

Layer	Node representations	
Input	Female features	Male features
	Dark skin	Light skin
	Lab coat	Tie
	Dress shoes	Gang headband
	Sweatpants	Gang tattoos
	Suit jacket	Street Signs
	Cars	Hospital beds
	Nursing station	Receptionist
	Cubicles	
	Context	Street corner
Professional clothing		Athletic clothing
Gang clothing		Hospital
Profession	Businessperson	Athlete
	Doctor	Gang member
Attributes	Caring	Athletic
	Popular	Rich
	Greedy	Violent
	Intelligent	Unintelligent

consist of two nodes, one layer representing either male or female, and the other White or Black. Race and gender are represented in this network by single nodes. (It would have been possible to build the current network so that it learned to identify race and gender from more basic physical cues, but this increased complexity would not have influenced the ability of the network to simulate our key points.) The race and gender conjunction layer consists of eight nodes and learns to represent conjunctions of race and gender: Black male, White male, Black female, and White female.

Higher order semantic knowledge was represented by five layers: (a) the context layer, with 3 nodes that represented the individual's physical context: street corner, office building, or hospital, and 3 that represented the individual's clothing: professional clothing, athletic clothing, or gang clothing; (b) the profession layer, with 4 nodes that represented 1 of 4 possible professions for the observed individual: doctor, gang member, business person, or athlete; (c) the attribute layer, with 8 nodes that represented possible personality/social attributes of the individual: caring, athletic, popular, rich, greedy, violent, unintelligent, and/or intelligent; and (d) two hidden layers, each consisting of 100 nodes. The context, profession, and attributes layers were localist layers, with each node representing a specific item (e.g., street, doctor, intelligent). Hidden Layer 1 enables the network to learn combinations or conjunctions of input features that are associated with specific contexts, professions, and personality/social attributes. Hidden Layer 2 enables the network to learn combinations of situation and clothing that are associated with different professions.

For each of the layers, we set a level of inhibition within the layer that would control how many nodes would remain active after processing. Inhibition for the race, gender, and race and gender conjunction layers was set to strongly favor

only the activation of one node. We used stricter inhibition for profession because only the most strongly activated profession should be active, whereas we used more lenient inhibition for the attributes, because the number of possible attributes varied for different individuals. The context layer utilized inhibition that drove the layer to prefer to activate only two or so nodes at a time. Hidden Layers 1 and 2 utilized inhibition that drove the layer to activate around 20% and 25%, respectively, of the nodes at one time (see Appendix A for more details).

Given research indicating that positive and negative evaluations are the result of two distinct processes (Cacioppo & Berntson, 1994; Cacioppo et al., 1997), both a positive and negative evaluation layer were included as scalar value layers. The non-linear activation functions that are typical for most neural network models do a poor job of representing relatively linear, continuous values, as they have a strong binary output bias (0, not activated, to 1, activated). However, relatively linear values can be coded across a set of nodes within a layer. The two evaluation layers were constructed of 12 nodes to produce scalar output values that range from 0 to 1.0 with higher values indicating a more positive or negative evaluation respectively. Although it should be noted these layers can occasionally exhibit values that are slightly out of this range.

**Connections.** The network contains both unidirectional and bidirectional connections (see Figure 2). Two sets of parameters were used in the network to specify the learning rate and the proportion of Hebbian and error-correcting learning. Learning details are presented in Appendix B.

The Leabra architecture uses two forms of learning, Hebbian or associative learning, and error-correcting learning. Hebbian style learning is sensitive to the covariation among features, whereas error-correcting learning is sensitive to whether the network correctly predicts or classifies the category or features of an object or event. Although some might think that these two forms of learning might correspond to dual processes in models of attitude formation, they do not. First, both work by modifying the weights among nodes, and they occur at the same time, as learning in Leabra is implemented by averaging together the weight change estimated by the two types of learning and using that average to modify the weights in the network. Second, error-correcting learning does not require consciousness or controlled processing and is found at multiple levels within the brain. Third, error-correcting learning is associative and not propositional, as it works by modifying the weights among nodes. Finally, the idea of error of prediction is not the same as validity or truth value, as it can apply to motor learning or reward learning, where we are not concerned with the truth of a proposition.

## Training

Training was completed in two steps, early and late training. Early training taught the network to recognize and evaluate the

highly salient perceptual cues of race and gender. Previous research has shown that attention is preferentially directed to both an individual's race and gender within about 100 ms to 150 ms after stimulus onset (Ito & Urland, 2003). Thus, the network was designed to quickly identify and process race and gender stimuli. Early training was utilized to establish stereotypes based on early perceptual analysis of race and gender cues. Late training allowed the learning of appropriate higher-level semantic representations of the stimuli and appropriate corresponding evaluations. Current training procedures are not designed to represent real life learning processes; instead, they are implemented simply to construct a network that had the hypothesized representations and knowledge.

**Early training.** During early training, only the input layer, gender, race, race and gender conjunction layers, and positive and negative evaluation layers were active. All other layers were lesioned (lesioned layers do not send or receive any information). This allowed the network to learn perceptually based evaluations of race and gender that were not influenced by higher-level semantic systems. For early training, four sets of stimuli were presented representing a White and a Black individual of each gender along with the corresponding positive and negative evaluations.

To examine the impact of differences in the strength of these stereotypes on processing, we performed three separate sets of simulations using three different sets of early training stimuli that systematically varied in the strength of the negative evaluations for Black males and Black females, with strong, moderate, and slight negative Black stereotypes. The early training evaluations for White males and females were consistent across all early training instances.

For early training, in the input layer, only stimuli representing racial and gender-specific features were presented. For the evaluation layers, the specific evaluation was presented. Early training evaluations were driven by perceptually based race and gender stereotypes. The evaluations we used are designed to be illustrative of the impact of different evaluations for different kinds of individuals. We are not arguing that they are precisely representative of actual differences within social perceivers. Black male and female each had a positive evaluation of 0 across all three sets of early training stimuli (i.e., no positive evaluation). However, Black male and female negative evaluations were systematically varied across the three training sets: .8 and .5 respectively in one set, .5 and .3 in a second set, and .2 and .1 in a third set (i.e., varying degrees of negative evaluation). White male and female each had a positive evaluation of 1.0 and .9 respectively and both had a negative evaluation of 0 (i.e., highly positive evaluations and no negative evaluations for Whites). Thus, in addition to being racist, the network was trained to be mildly sexist, evaluating White males slightly more positively than White females. We intentionally chose to train the network to hold negative racial and gender stereotypes to examine the evolutions of attitudes about specific



individuals (e.g., Black male doctor) that would not be present in people not holding these stereotypes. As such, the network is not meant to represent a “typical” individual. There was no need to vary the training ratio of 1:1:1:1 to get different evaluations of the four types of individuals, because the scalar layers used for the evaluation layers allow for direct training of evaluation. The network performed 50 epochs of training, and the final weights learned were saved for use in late training.

**Late training.** Before late training was conducted, weights were loaded from early training and all layers were unlesioned so that all layers in the network now processed activation. The network performed 200 epochs of late training, for a total of 250 epochs of training across both training steps. More epochs of training were required for later training because the relationships in later training were more complicated than the simple associations acquired in early training.

For late training, multiple sets of stimuli were presented with varying frequencies (presented in Table 2) in a random sequence representing race and gender-specific doctors, business people, athletes, and gang members. Each of these “individuals” was represented across all 17 nodes of the input layer, with nodes representing racial and gender features, appearance features such as clothing that were diagnostic of different professions or roles, and environmental features that were diagnostic of different physical contexts (see Table 1). The network was trained to use these 17 input features to recognize one of four possible professions, one of three physical locations, one of three types of clothing, and various different personal attributes (see Table 3). During training, a set of input features was applied to the input layer and appropriate activations were applied to the target layers so that the network would learn to associate the correct input features with the desired profession, context, and attributes. For example, in learning about a Black female doctor, the input features that identified this individual and her context (dark skin, female features, lab coat, tie, dress shoes, hospital beds, nursing station, receptionist) would be set at 1 and the desired target values (professional clothing, hospital, doctor, caring, popular, rich, intelligent) would be set at 1 in the appropriate target or output layers.

Evaluations resulting from iterative processing of higher-order semantic layers were based on the profession and corresponding attributes of an individual and did not vary by race or gender (e.g., gang members were highly negative, doctors highly positive regardless of race or gender). Doctors had a positive evaluation of 1.0 and a negative evaluation of 0. Businesspeople had a positive and negative evaluation of .5. Athletes had a positive evaluation of .6 and a negative evaluation of .2. Gang members had a positive evaluation of 0 and a negative evaluation of 1.0. In addition, there were instances of just Black or White males or females, as in early training where only racial and gender stimuli were presented, and the respective evaluations were the same as in early training.

**Table 2.** Early and Late Training Individual Presentation Frequencies.

Individual	Presentations per epoch	
	Early training	Late training
Black male	1	3
Black female	1	2
White male	1	3
White female	1	2
Black male doctor		3
Black male businessman		2
Black male athlete		6
Black male gang member		12
Black female doctor		2
Black female businesswoman		1
Black female athlete		4
Black female gang member		3
White male doctor		6
White male businessman		6
White male athlete		6
White male gang member		6
White female doctor		6
White female businesswoman		4
White female athlete		5
White female gang member		4

Note. For training with no Black male doctors, all frequencies remain the same except no Black male doctors are presented.

**Table 3.** Assigned Inputs, Contexts, Professions, and Attributes.

Profession (learned)	Input	Context (learned)	Attributes (learned)
Doctor	Lab coat, tie, dress shoes, patient beds, nurse station, receptionist	Professional clothing, hospital	Caring, popular, rich, intelligent
Businessperson	Tie, dress shoes, suit jacket, receptionist, cubicles	Professional clothing, office building	Rich, greedy, intelligent
Athlete	Sweatpants, street signs, cars	Street corner, athletic clothing	Athletic, popular, rich
Gang member	Gang headband, sweatpants, gang tattoos, street signs, cars	Street corner, gang clothing	Violent, unintelligent

Two late training contexts were utilized in two separate sets of simulations. The primary simulation was done to test the primary predictions about the evaluation time course for different individuals and to see how the time course from earlier and later evaluations evolve. This context and simulation included all individuals and their frequency of presentation, which is presented in Table 2.

A follow-up simulation was done, with the second late training context, to test whether the evaluation of a novel,

never seen before individual (i.e., a Black doctor) depended on actual experience with that specific individual or whether the network could integrate information about race and gender stereotypes with general information about other (non-Black, male) doctors to arrive at a positive evaluation of a Black doctor. Although this simulation focuses on the Black doctor, we are simply using this individual to test the general process of integration. The second learning context did not include Black male doctors; all other frequencies remained the same.

### Testing

For the basic simulations as well as the first and second follow-up simulations, all input features were presented simultaneously. The network was then allowed to process the stimuli for 80 cycles. The evaluation time courses generally settled on a final evaluation by Cycle 60; thus, more iterations would not result in any further evaluation changes.

For the third follow-up simulation (i.e., activation of context prior to full event activation), just features specific to the doctor profession (situation and clothing, but no race or gender information) were presented initially (see Table 3 for context stimuli), and the network was allowed 30 cycles to evaluate the context before the full event stimulus (i.e., context features plus race and gender features) was presented and the network was allowed to continue settling. The decay proportion between context and full event stimulus was set to .30, so that 70% of the activation due to context was still present, when the full event stimulus was presented. This allowed the network to simulate evaluating an individual's context immediately before evaluating the individual him or herself to determine the effect of prior activation of context on the evaluation time course of an individual.

To ensure that the results were not specific to a single run, for all simulations, we performed 10 different training and testing trials and then averaged the results across the 10 runs. Between each run, consisting of a training and testing trial, the network was reinitialized, assigning different random, weak starting weights, thus ensuring that each training and testing cycle would start from a different set of random weights and activations.

### Results

The dependent variable of interest was the time course of both positive and negative evaluation. We include only graphs most relevant to the predictions in the main part of the article; but all graphs can be found in Appendix C. For all simulations, we tested three different levels of stereotype strength (i.e., strong, moderate, and slight negative Black stereotype) as established in early training. We crossed this with two late training contexts, one with all individuals (primary simulation and first follow-up) and one with no Black male doctors (second follow-up simulation). In addition, in a third follow-up, we tested the impact of presenting context

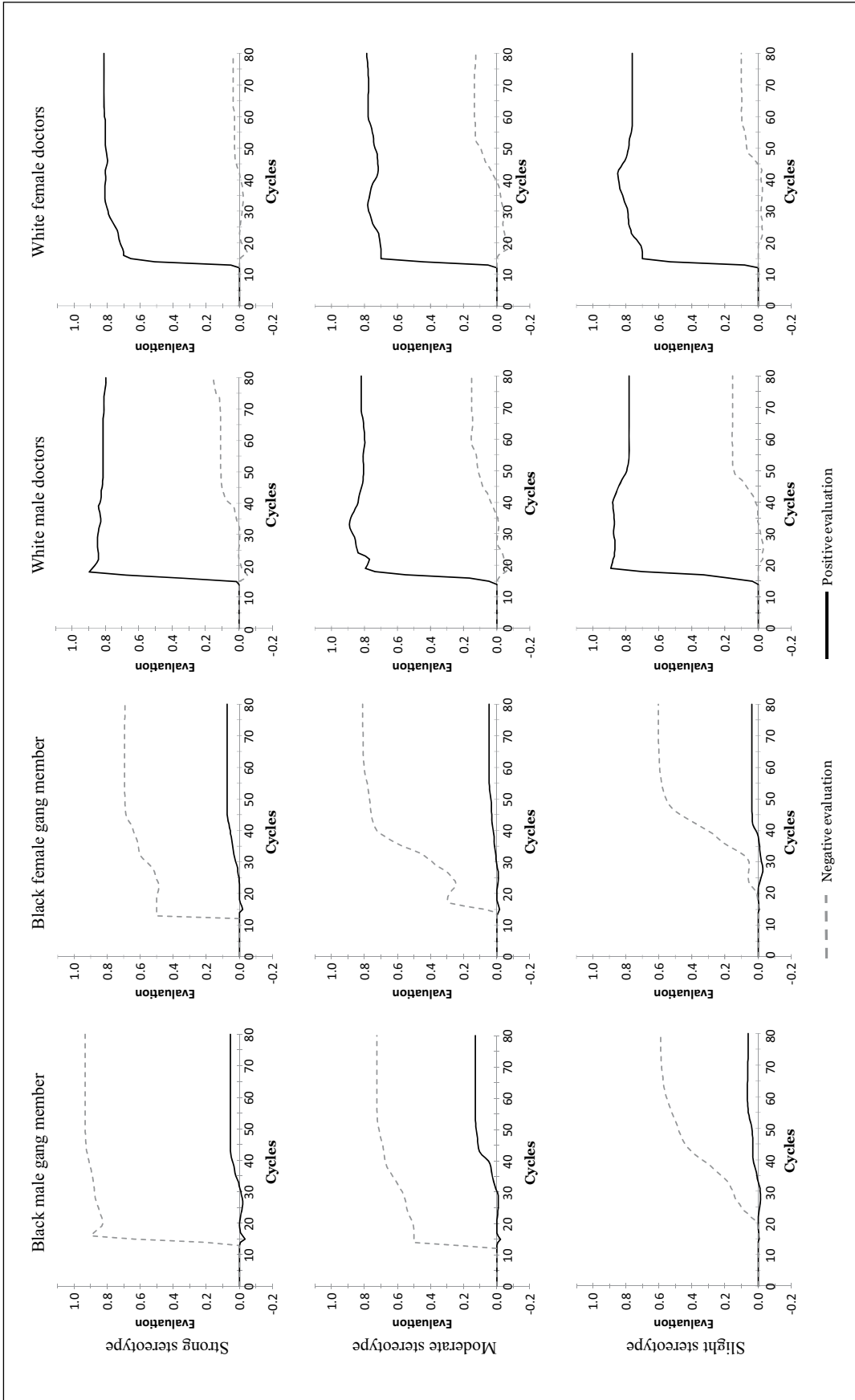
information *before* receiving person information, on the time course of evaluation. This resulted in a total of seven different simulations. For each simulation, we performed 10 different training and testing trials and then averaged the results across the 10 runs for 20 individuals: a Black and White, female and male of each profession, as well as a Black female, Black male, White female, White male.

Below, we present the most relevant results testing the four main predictions: (a) Evaluations would dramatically shift over the time course of evaluation when earlier stereotype evaluations conflict with semantic information processed by later iterations of the network, and that there would be no shift in evaluations when earlier processed information is consistent with information processed in later iterations; (b) the strength of the earlier negative evaluations will decrease as the strength of the learned negative stereotype decreased (from strong to moderate to low); (c) a novel combination of stimuli never learned before (i.e., a Black male doctor when the network was never previously exposed to a Black male doctor) will be integrated and will show a similar pattern of evaluation to that exhibited when the stimuli had been previously encountered; and (d) the activation of a context immediately prior to race and gender cues will influence the evaluations of a specific individual. Specifically, in the case of the Black male doctor, when a positively evaluated context (hospital and doctor clothing) precedes information about the individual's race and gender, the negative stereotype based on race and gender will have little to no influence on the doctor's evaluation, even at the earliest time step.

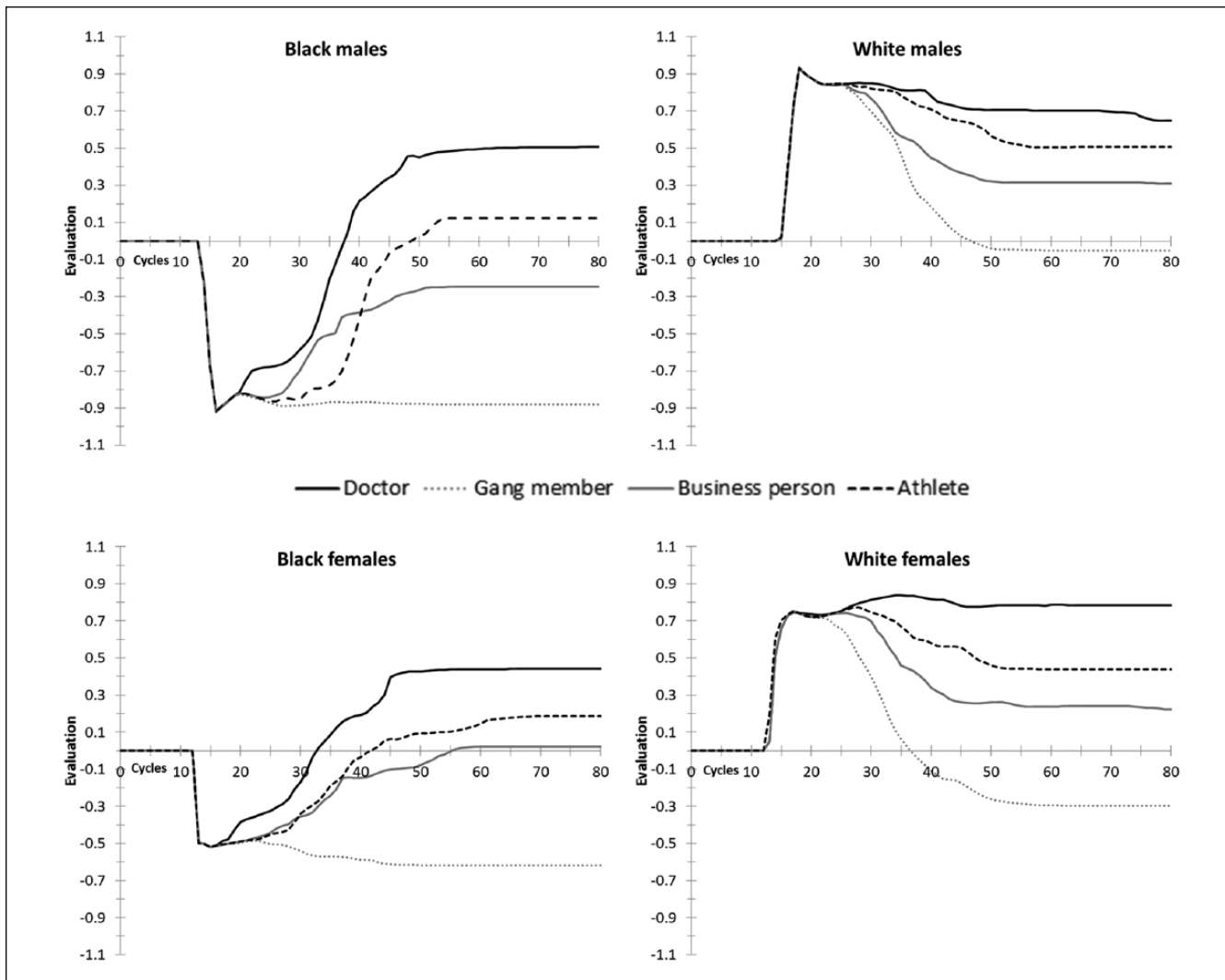
### Evaluation Time Courses

*Congruent earlier and later information.* As predicted, when earlier processing based on learned stereotypic information is congruent with later semantic information processed in further iterations (e.g., Black male gang members, White male doctors), no dramatic shift was seen over the evaluation time course (see Figure 3). In addition, to facilitate comparisons across different individuals, a net evaluation was calculated for each individual by subtracting the negative evaluation from the positive evaluation. This net evaluation time course for the four professions by each race and gender-specific individual for the strong stereotype training is graphically displayed in Figure 4 (see Appendices D and E for moderate and slight stereotype net evaluation graphs).

*Conflicting earlier and later information.* When presented with an individual whose earlier processed stereotypic information conflicts with semantic information activated by later processing, we expected a dramatic shift in evaluations. The critical test of this was a Black male doctor, an individual with the largest discrepancy between a negative learned Black racial stereotype and positive evaluations based on profession. Indeed, we observed exactly this pattern of shifting from a highly negative earlier evaluation to a highly positive later



**Figure 3.** Evaluation time courses of congruent earlier stereotypic and later information for Black gang members and White doctors.



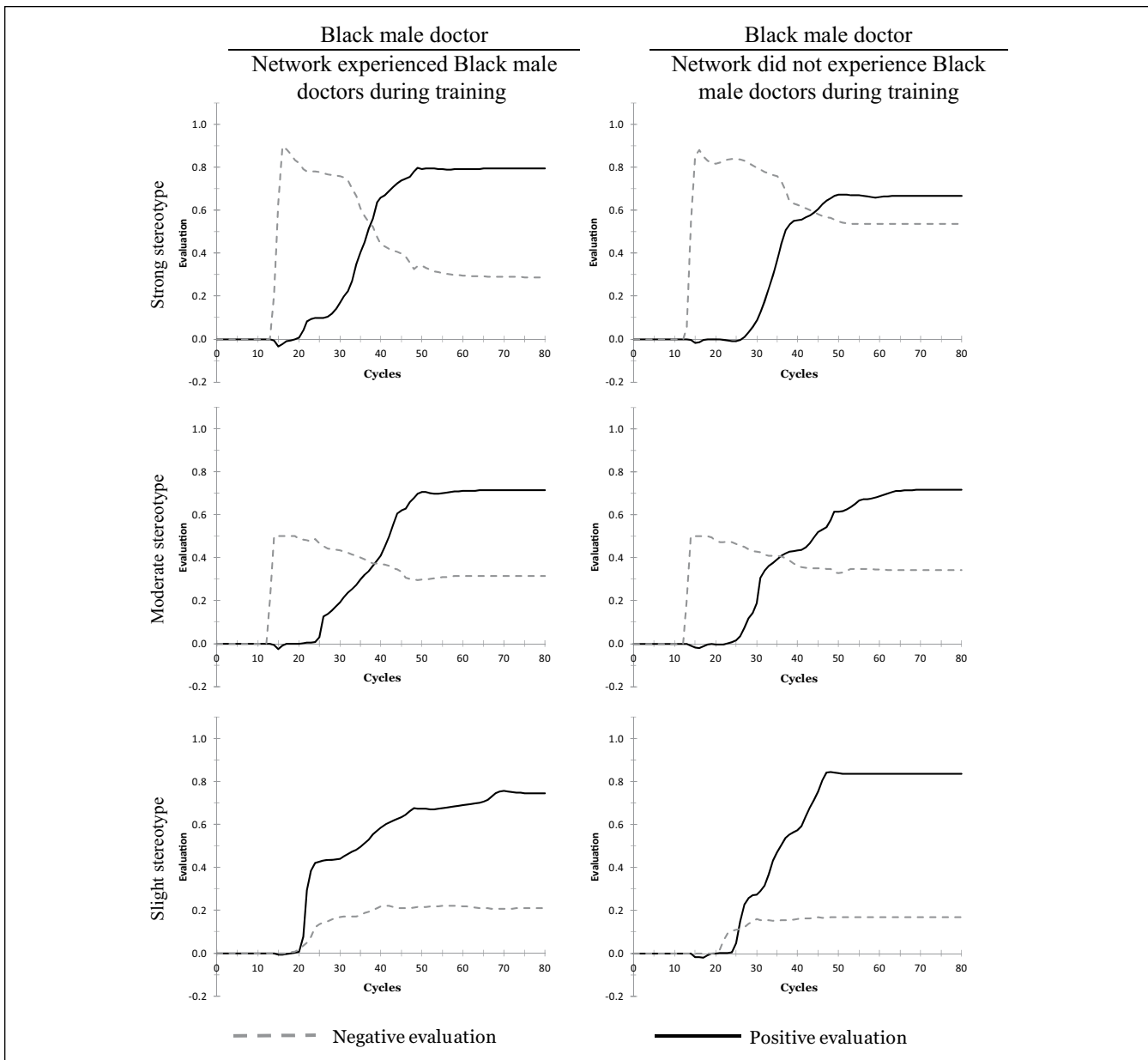
**Figure 4.** Net evaluation time courses for strong stereotypes for all professions, by race and gender.

evaluation, when the network learned a strong negative Black stereotype (see Figures 4 and 5). A similar pattern was observed with other instances of conflicting earlier and later information processing (e.g., White male gang member, Black male athlete; see Appendices C, D, and E). Among females, similar patterns to males are found, although the shifts were more moderate because of the smaller differences between evaluations of Black and White females.

*Influence of stereotype strength.* As expected, the weaker the learned negative Black stereotypes, the weaker the earlier negative evaluations of individuals. When evaluative information was congruent throughout the evaluation time course of processing (e.g., Black male gang member), the weaker the early stereotype, the weaker the impact of the negative stereotype on the earlier evaluation, compared to the impact of the later processed semantic information, particularly for Black individuals (see Figures 3, 4, and 5; see

also Appendices C, D, and E). Similarly, when information conflicted during the evaluation time course, the strength of the earlier evaluations was attenuated as the strength of the learned negative Black stereotype was decreased, allowing semantic information processed later in the evaluation process to more heavily influence later evaluations.

*Processing of a novel individual.* As in the second follow-up simulation we were primarily interested in how the network evaluated a novel “individual” (i.e., a Black male doctor), thus, only the results from Black male doctors are presented. There were no other major differences in evaluation patterns or later evaluations for the other individuals. For the strong stereotype conditions, there was a greater discrepancy between the later evaluations for the two late training contexts. When the network had not experienced a Black male doctor before, it exhibited a strong earlier negative evaluation that only moderately decreased over time while a fairly

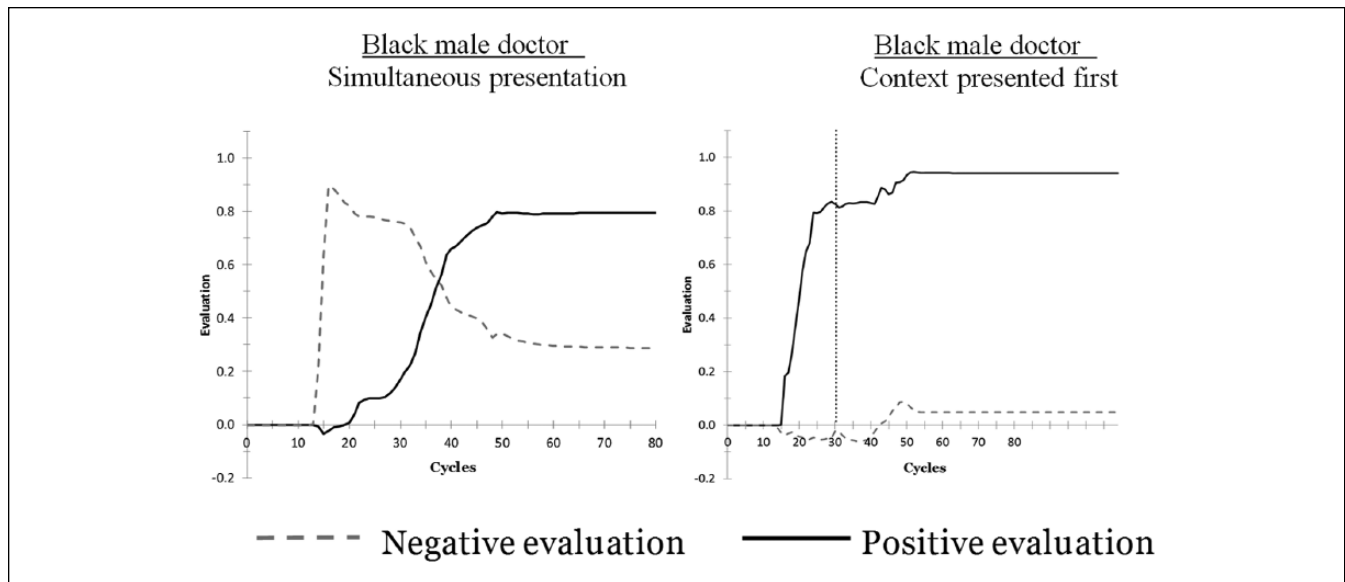


**Figure 5.** Evaluation time courses of conflicting early stereotypic and later information for Black male doctors, given late training with and without presentations of Black male doctors.

strong positive evaluation appeared later (see Figure 5). The network that had experienced Black male doctors previously also showed a strong earlier negative evaluation spike, but the negative evaluation decreased sooner, more rapidly, and settled on a much lower negative evaluation than the network that experienced a Black male doctor for the first time. Furthermore, for this network, the positive evaluation increased sooner, more rapidly, and settled at a higher final positive evaluation than for the network that experienced a Black male doctor for the first time. Regardless of whether the network had experienced a Black male doctor before, there were no major differences in the evaluation patterns and later

evaluations of Black male doctors for the slight and moderate stereotypes, with the exception that the increase in positive evaluation occurs later when evaluating a novel Black male doctor. Nevertheless, across all stereotype strengths, even when the network had not previously experienced a Black male doctor, it showed a similar pattern of changing from a negative to a positive evaluation over time.

*Prior activation of context.* In the third follow-up simulation, we were interested in determining the effect of processing context received immediately prior to receiving race and gender cues about a specific individual. Again, we chose to use



**Figure 6.** Comparison of evaluation time courses of Black male doctors with and without activation of context before presentation of the race and gender features.  
 Note. Dashed vertical line indicates when full event (i.e., addition of race and gender) cues were presented to the network when context was activated first.

the Black male doctor as the individual of interest to observe the effect of prior activation of context on the evaluation time course of the Black male doctor. When the environmental context (i.e., hospital) and clothing information (i.e., professional clothing) were presented before the full event stimulus (which includes additional information about race and gender), the network quickly and accurately evaluated the context very positively with no negative evaluation (see Figure 6). When the full event was then presented (i.e., all the stimuli for Black male doctor that was used in prior simulations), the network increased its positive evaluation of the individual over time, and there was a negligible increase in negative evaluation that decreased to nearly zero over time. In comparison with the simulation where the full set of stimuli for the Black male doctor was presented simultaneously, when the context was presented first, there was no strong earlier negative evaluation of the Black male doctor, but instead, the evaluation stayed quite positive (see Figure 6).

## Discussion

The network supports the theoretical IR model by demonstrating continuous evaluation modification over time in an integrated system comprised of multiple processes. The traditional understanding of evaluation in social psychology is based on dual process and dual system models that propose two distinct evaluation systems or processes: an implicit and an explicit process/system. The theoretical IR model provides an alternative understanding of evaluation that proposes an evaluation processing circuit that integrates across multiple neural processes to produce evaluations that evolve over time. Importantly,

we do not mean to suggest that the IR model is more parsimonious than other existing models of evaluation; instead, we believe that the theoretical IR model more accurately represents the neurological construction and operations underlying evaluation. Consistent with the theoretical model and supporting the four main predictions, the neural network constructs evaluations from the input of multiple perceptual and semantic systems that are determined by a set of factors such as context, learned stereotypes, and profession specific attributes.

## Evaluations Evolve Over Time

In the simulations, each individual presented to the network was quickly evaluated based on stereotypical racial and gender stereotypes. Evaluations then continued to evolve as further iterations recruited higher-level semantic layers as represented by the context, profession, and attribute layers that then influenced the positive and negative evaluations for the individual presented to the network. For example, the White male doctor and Black male gang member (Figure 3) recruited further attributes that were supportive of the network's stereotype, so further iterations over time only strengthened and/or maintained the network's earlier evaluations. However, further processing beyond earlier evaluations for the Black male doctor (Figure 5) activated characteristics and related attributes contradictory to the earlier processed stereotype information. When these later continuing iterations processed contradictory characteristics, the evaluations evolved in a drastic manner. In the case of the Black male doctor, the context of a hospital, recognition of "doctor" as the individual's profession, and processing of attributes such as intelligent,

popular, rich, and caring resulted in an increase in positive evaluation and a decrease in negative evaluation that revised the network's earlier evaluation, thus resulting in an evaluation time course that changed dramatically over time.

Similar, but more nuanced results were found for females. Although the network was trained with a negative stereotypic evaluation for Black females, this stereotype was not as strongly negative as for Black males. The gender differences in earlier evaluation resulted in lower negative evaluations for females and higher positive evaluations when compared with the relevant Black male counterpart as would be expected with less robust negative stereotypes for both the strong and moderate stereotypes. In addition, Black females generally settled on a slightly more positive and slightly less negative evaluation, and White females were slightly less positive than White males in the initial training. This gender difference led to White females generally being evaluated more negatively and less positively over the evaluation time course than their White male counterparts for all professions. The one exception to this was the White female gang member. White female gang members had a higher final negative evaluation than White male gang members in the strong Black stereotype condition, as expected, but lower final negative evaluations along with more positive later evaluations than the corresponding White male gang members in the moderate and slight stereotype conditions. Although the discrepancies in later evaluations across conditions were unexpected, as there was no manipulation of initial White evaluations between stereotype strength conditions, we suspect that these results were likely due to the fact that the network had difficulty learning appropriate evaluations for White female gang members driven by the fact that the network was presented with (and thus could learn from) fewer female gang members than male gang members (7 vs. 15) and also that the White female gang members had more discrepant earlier and later evaluations than the Black female gang members, making learning appropriate evaluations more difficult.

### *Impact of Different Strengths of Initial Black Stereotype*

The results demonstrate that negative Black stereotypes of different strengths resulted in earlier evaluations that were consistent with the strength of the stereotype. As the network completed further iterations of the inputs and further processing of higher level concepts such as the individuals' contexts, professions, and attributes, the evaluations changed over time. Importantly, it is evident that these earlier stereotypic evaluations can be largely overcome by later information, although the later evaluations were still influenced by earlier reactions (e.g., Black male doctors and White male doctors never have equal later evaluations).

The evolution of the evaluations also has theoretical implications. A dual process or dual system interpretation of the results would attribute the earlier evaluations to an

implicit process or system and the later evaluations to an explicit process or system. However, consistent with the theoretical IR model, the results understood in the context of the current network demonstrate that a multiple process network can produce discrepant earlier and later evaluations (see also Monroe & Read, 2008).

### *Evaluations of Black Male Doctor When Not Previously Encountered*

Whether or not the network had ever encountered a Black male doctor did have some impact on his evaluation, especially when the network held a strong negative Black stereotype. When the earlier evaluation of the basic stereotype was only moderately or slightly negative, if a Black male doctor had not been encountered during training, the evaluation time course of the Black male doctor was quite similar to his evaluation time course when a Black male doctor had been encountered during training. However, when the network held a strong stereotype, the evaluation time course when the Black male doctor had not been encountered was different. It took more iterations for semantic information about the doctor to overcome the earlier negative stereotypic evaluation, and the later evaluation of the Black male doctor was more negative and less positive than when a Black male doctor was encountered during training, although it was still positive.

Overall, this pattern of findings suggests that seeing an individual in an unexpected positive role can lead to an evaluation that is quite different from one's earlier impression. Although a Black male doctor had never been previously encountered, the network could combine the doctor information with the Black male information and arrive at an evaluation that integrated the two sources of information. The new information could override the earlier evaluation particularly well when the earlier impression was only slightly or moderately negative. However, as the results in the strong stereotype condition suggest, if the earlier processed stereotypic information is highly negative, even highly positive novel information may be unable to totally override the earlier negative impression.

### *Evaluations When Context Was Presented Before the Individual*

When contextual features (i.e., situational context and clothing) were presented before race and gender features for the Black male doctor, we observed that the earlier spike of negative evaluation was greatly inhibited (Figure 6). This finding carries significant implications, as it shows that very early attitudes of individuals can indeed be influenced by contextual factors present before the specific individual is evaluated. This finding is supportive of the growing literature demonstrating the early attitudes are indeed influenced by factors such as context (e.g., Barden et al., 2004). As

previously discussed, this finding is hard to handle for many dual process and dual system theories that propose that implicit or automatic attitudes (i.e., very early attitudes) are recalled object–attitude association or require that later controlled processing influences or corrects implicit or automatic attitudes. In addition to context, recent research is emerging that suggests that additional factors such as motivation can also influence very early evaluations (Cunningham et al., 2012). Although we did not include motivational processes in this neural network, we do not believe that incorporating motivation or additional factors such as emotion would pose serious difficulties; although it would require a more advanced network. As research continues to grow our understanding in this area, we believe a more exhaustive network including more factors would be very useful.

### *Implications for Attitudes and Attitude Processes*

The patterns of evaluation in the current simulations are consistent with patterns that are commonly attributed to separate implicit and explicit evaluation processes or systems. However, the current network demonstrates that we do not need to postulate two independent systems, processes, or attitudes, but instead we can postulate a dynamical evaluation that quickly evolves over time as additional neural systems are brought online and information is processed.

The connectionist neural network furthers our understanding of the underlying mechanisms by shifting our attention away from the idea of distinct implicit and explicit processes or systems and focusing it on the difference between earlier and later iterative processing. Earlier processing depends on quick perceptual processing and results in quick and automatic evaluations whereas later iterations recruit a dynamic interaction between several bottom-up and top-down processes that allow attitudes to be established in accordance with current contexts and goals. This development of evaluations elucidates important implications in our understanding of evaluation as well as the nature and establishment of attitudes.

*Are attitudes stored or constructed?* This neural network has significant implications for the stability of attitudes and for the issue of whether attitudes are stored or constructed. The network suggests that attitudes are stored in a given set of weights among nodes, so that when nodes are activated, the activations combine to create an evaluation. In this network, the *associations* between evaluation and objects and attributes are stored and relatively stable, although they can change with learning. However, the evaluation of a particular person or object depends on changing patterns of activation; it is dynamic and develops over time and is strongly influenced by context. Thus, the network shows how a stable evaluative representation can nevertheless result in variable and dynamic attitudes. As we demonstrated with the example of a Black male doctor in the primary simulation, the

network captures situations in which an evaluation of an individual can change dramatically over time. In this specific case, the network is showing how an initial, basic stereotype may function, and more importantly, how it can be overridden by further processing.

*How many attitudes are there?* An additional implication of the network is that the potential number of attitudes a person can hold is large, if not infinite, as evaluations are constructed based on the extent of evaluation processing and available cues. Evaluations develop over time in response to the attributes of the individual or object and the attributes of the context they are in. Thus, there are multiple potential attitudes, and it is rarely possible, a priori, to specify how many there are.

*What is the “true” attitude?* One reason that researchers have been so interested in implicit attitudes is the sense that they represent an individual’s “true” attitude. However, consider the example of the Black doctor in the primary simulation. Is the early negative response the “true” attitude? This does not seem to be necessarily so. The earlier attitude and the later attitude are based on very different brain systems and different bodies of information. The earlier attitude is based on a quick initial, almost perceptual, characterization of the individual as a Black male, whereas the later characterization is based on a much richer representation that includes information from the context and his profession, as well as possible encounters with Black doctors. In the example of the doctor, we would argue that the later evaluation is the “true” one as it best represents what the perceiver really feels about the target.

Furthermore, we can see that simply presenting information regarding an individual’s context immediately prior to race and gender information can dramatically alter the earlier evaluations of that individual, further supporting the notion that early “implicit” attitudes are not invariant and are likely not a person’s “true” attitude.

The current network, by presenting an explicit account of how information is recruited and processed over time, makes it much clearer that attitudes can differ very strongly depending on the patterns of attributes on which they are based. This particular argument is not a particularly novel one, especially from a connectionist perspective, which views evaluation as a process that develops over time as activation spreads through a network representing numerous potential attributes of an individual (e.g., Bassili & Brown, 2005; Conrey & Smith, 2007; Monroe & Read, 2008; Read & Monroe, 2009).

*How many processes are there: One or two?* Dual process and dual system models have led to many insights into attitudes and attitude formation. However, we propose that continuing research and theorizing should expand beyond the constraints of fitting the evidence and theories to two (or one) processes or systems, especially two distinct and independent processes or systems. Rather than trying to make the case for a fixed set



of processes or systems, either 1 or 2, and in light of new evidence revealed by technological advances (e.g., Freeman & Ambady, 2011; Van Bavel et al., 2013) and from the presented neural network, we should instead be investigating the relevant mechanisms and trying to understand how the many different processes or systems in the brain are involved in forming our attitudes.

These sentiments are echoed by others, such as Evans (2008), who points out that it does not really make sense to talk as if there is a single System 1 or impulsive process. Within System 1, we have systems responsible for language, vision, audition, and touch, and within these broad systems, we have further subsystems that are responsible for multiple different components. For example, with language, we have systems responsible for semantics, phonology, syntax, and visual word recognition. Each has been treated as a different system with its own representations and computations. Treating all of System 1 as one large system may hide important processes and thus limit our ability to understand how the mind works.

*Implicit versus explicit attitudes.* This network and the simulations are also relevant to understanding the distinction between implicit and explicit attitudes, one of the most active areas of research in the past 10 to 15 years in social psychology. Implicit attitudes are argued to be quick and automatic, and involve little or no conscious control or regulation, whereas explicit attitudes occur later and are more deliberative. As various researchers have noted, early research tended to confound the type of measurement with the type of construct actually being measured. It has been suggested by De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009) that it may make more sense to refer to indirect and direct *measurement* and implicit versus explicit *evaluations or attitudes*.

Typical implicit and explicit measures differ along several dimensions. First, implicit measures typically measure evaluation as a “side effect” of something else, such as priming and reaction time, or word-fragment completion. The respondent is not asked to deliberately give his or her evaluation. In contrast, explicit measures ask respondents to deliberately give their evaluation. Second, implicit and explicit measures typically have different time courses. Implicit measures typically measure an evaluation or attitude shortly after it is activated by a stimulus (although see Sekaquaptewa, Vargas, & von Hippel, 2010, for pencil and paper implicit measures), whereas explicit measures have a longer time course and can measure an evaluation after more processing time has elapsed. Thus, the attitude measured by an explicit measure may be the result of much richer and more detailed processing. Third, implicit and explicit measures differ in terms of the extent to which deliberative or executive function processes are involved in the evaluation. Implicit measures provide less opportunity for the influence of executive function (although they do not eliminate its impact), whereas explicit measures allow for extensive influence by executive processes. As a result, explicit measures may be much more strongly influenced by motivations, such

as self-presentation or desire to appear non-prejudiced. Thus, attitudes measured by implicit and explicit measures may differ for at least three different reasons.

Dual process models tend to equate reflective or controlled processing with later developing processes. However, the two components can be somewhat independent of each other. As the current network demonstrates, an evaluation can develop over time through the iterative reprocessing of a stimulus, without necessarily involving reflective or controlled processing. And as the current network has shown, the later developing attitude can be quite different from the earlier attitude, without the involvement of any kind of controlled processing.

*Associative versus propositional processes.* Some theorists such as Strack and Deutsch (2004) as well as Gawronski and Bodenhausen (2006) argue that central to the distinction between implicit versus explicit processes (impulsive vs. reflective processes) is the distinction between associative and propositional processes. According to these authors, associative processes rely on representations based on similarity as well as contiguity, and processing occurs through the spread of activation. Whereas propositional processing has a language-like form that is manipulated through logical, explicit reasoning processes. Moreover, the results of propositional processing, unlike associative representations, have a truth value. They argue that implicit processes rely on associative processing, whereas later, more reflective (explicit) processes rely on propositional processes. They argue that a major reason that implicit and explicit attitudes may be different is that these two different processing systems may result in different information or attitudes. That is, the results of the associative processing of a stimulus and the propositional processing of a stimulus may be quite different.

However, the current network provides an account for how implicit and explicit attitudes may often differ that does not rely on that distinction. All the processes and representations in the current network are what they would call associative, yet we can see that the earlier evaluation of a stimulus may differ fairly dramatically from its later evaluation, without assuming different forms of representation and process.

*Unimodel.* In contrast to various dual process and dual system models, Kruglanski has argued for what he terms a unimodel (Kruglanski & Gigerenzer, 2011). He proposes that all information processing can be characterized as making inferences from various kinds of evidence. Moreover, all of this processing is rule based, and what looks like different processes are instead the use of different rules with different parameters. Thus, there is only one process, namely, rule-based processing. As a result, dual process models that depend on the distinction between associative and rule-based processing are not compatible with the unimodel’s single rule-based process.

However, this argument assumes that the only basis on which to postulate different cognitive processes is whether

they are rule based or not. However, this may be problematic, as the unimodel uses a meaning for process with which most psychologists would probably disagree. For example, let us take the idea of a connectionist or neural network model of the brain. In such models, everything occurs in terms of the passage of activation among nodes along weighted links. Essentially, the model relies on associative processes. So in Kruglanski's use of the term *process*, a connectionist model of the brain is a unimodel. In his terms, vision, audition, and social perception all use the same process. At one level, this may make sense, as it is useful to point out the common substrates and mechanisms of how the brain operates. However, at another level, it is insufficient. Vision, audition, and social perception rely on different brain systems and use different representations and different computations. A potentially more useful model of social perception would need to go beyond stating that perception relies on neurons and associative processes, and instead would outline how those neurons and associative processes make up representations, and what the specific computations are that they perform that are central to social perception. A similar argument can be made for any claims about rule-based processes.

### *The Value of Dual Process and Dual System Theories*

Although we propose, based on theoretical and empirical evidence (see Cunningham et al., 2007; Van Bavel et al., 2012; Van Bavel et al., 2013), and supported by the current neural network, that we can understand attitude formation without proposing two distinct processes or systems, we are not arguing that there is no value in dual process and dual system theories. Dual process and system theories have led to a wealth of knowledge about attitudes and attitude formation. For example, understanding that we can have very early attitudes that vary greatly from our later attitudes, often without our awareness, is an important finding with many real world implications (e.g., implicit racism). Our argument for the need for a more dynamical understanding of attitude formation does not invalidate the findings or advances provided by the many dual process and dual system theories, and indeed, it may be that dual process and dual system models are moving to a more dynamic understanding of social cognition.

We, however, want to propose more than just dynamic interactions between two systems, and to support a more recent theoretical model, the IR model, as a means to continue to expand our understanding of attitudes. And we do so by expanding and building on the wealth of knowledge generated by the dual process and dual system theories. After all, there would be no interest in a dynamical understanding of attitudes if dual process and dual system models did not show that we can in fact have two distinct attitudes toward a single object! Nevertheless, we have tried to highlight in the discussion important implications of a more dynamical perspective for the understanding of attitudes.

### *From a Neural Network to Human Subjects*

For the simulations, we made four general predictions that were supported by the behavior of the neural network. Here we discuss how future empirical work can explore each of these predictions with human subjects, providing additional insights into the dynamics of evaluation. The first prediction stated that when initial stereotypic information is inconsistent with the evaluative implications of later activated information, then the evaluation will change over time. New methodologies may allow us to track this dynamic change in evaluation. Hand movement tracking is one potential research methodology that is designed to measure the real-time unfolding of underlying cognitive processing by tracking the manual action of participants as they indicate an evaluation or judgment (i.e., tracking the trajectory of a mouse cursor; see Freeman & Ambady, 2010; Freeman, Dale, & Farmer, 2011; Wojnowicz et al., 2009). In addition, EEG provides high temporal resolution measurement of brain activity, and advances in measuring event-related potentials (ERPs) from EEG may provide a means to measure the dynamics of evaluation over time (Amodio, Bartholow, & Ito, 2014). However, EEG is limited in its ability to measure the source of brain activity, but new simultaneous EEG and fMRI methods may overcome this limitation (for simultaneous EEG-fMRI methodology, see Huster, Debener, Eichele, & Herrmann, 2012).

The second prediction posited that the strength of a stereotype, operationalized as the extremity of the associated valence, will influence the early evaluation of a target, as well as how it changes over time; the stronger the stereotype, the more it will influence the early evaluation and its evolution. The current network explicitly tested the strength of stereotype valence (from slight to strongly negative). Another aspect of stereotype strength is based on the frequency of exposure to the stereotype or how well learned it is. Although not directly examined with the current network, it should be the case that the more frequent the exposure to a negatively stereotyped group or individual (with valence held constant), the stronger the impact of the stereotype on early evaluation and how it changes. Thus, experimental work could manipulate stereotype strength, both extremity of valence and frequency of exposure, and measure the effect of stereotype strength on the dynamics of evaluation.

Similarly, experimental work could examine the third prediction that individuals can successfully integrate information regarding a novel individual with an existing stereotype and produce a dynamic change in evaluation, such as the one we demonstrated with the Black male doctor example. Although a different target would need to be used (i.e., it is unrealistic to find participants for whom Black male doctors are truly novel), experimental work could establish stereotypes and then present novel individuals to participants and measure the time course of evaluation and also how the strength of the learned stereotypes influences the evolution of the evaluation over time.

Finally, the fourth prediction stated that the activation of context information prior to race and gender cues can override the impact of those cues. However, the temporal relationship between the context cues and the stereotype cues may have a strong influence on the particular pattern of evaluation because it influences which cues and evaluations are likely to be activated at which times. For example, when context is activated well before the stereotype cues are encountered, semantic and evaluative information may be strongly activated well before the stereotype cues are activated and thus, can override them. However, the closer in time that context gets to the reception of the stereotype cues, the less opportunity there is for the information activated by context to override the stereotype cues. If context and stereotype cues are received at the same time, as in most of the current simulations, then the stereotype cues should activate an initial activation that is only later overridden by the context cues. Thus, we predict that not only the temporal order of context matters, but also the temporal separation of context cues and other evaluative cues. Although it is not possible from the current network to specify how much time is needed for context cues to potentially override inconsistent stereotypic information, future work with the methodologies suggested above may be able to systematically investigate the influence of the temporal ordering of context cues.

### Limitations

The current network does not address one important aspect of the IR model. Cunningham and colleagues (2007) suggested that depending on such factors as motivation to process, or awareness of a conflict between two aspects of an earlier impression, the perceiver might process more extensively. For example, if the perceiver is aware that his or her earlier negative response to the doctor conflicts with their belief that he or she is not prejudiced, he or she may also think more extensively about the target, as well as invoke various kinds of self-control processes, such as self-presentation or suppression of prejudiced feelings. Future versions of the network should attempt to capture the role of motivation to process and the impact that awareness of a conflict might have on the extent of processing.

Another limitation of the network is that we do not test it against alternative computational models. This is because no computational implementations of such alternatives exist as far as we know. The closest is probably Freeman and Ambady's (2011) neural network model of person construal. However, that model focuses on categorization and not evaluation, and is more limited in the range of information it can handle. To construct a version of the Freeman and Ambady model that could be compared with the presented IR model, the revised Freeman and Ambady model would need additional layers: a context/situation layer, a layer that can infer profession, an attributes layer, and evaluation layers. With these additions, it is not clear that the revised Freeman and

Ambady model would be sufficiently distinct from our proposed IR model.

Furthermore, we do not know of any computational implementations of a dual process model, and there are several significant challenges to constructing convincing and fair comparison models. First, none of the dual process models of which we are aware are specified in enough detail to be directly translated into a computational model. We would have to make a number of assumptions about the authors' meanings and intentions that might result in a theoretically weak model. Second, despite our best intentions, constructing the models or networks we intend to argue against may not be as convincing as comparing the current IR model against alternative models constructed by those most strongly advocating for other theoretical approaches.

### Conclusion

This neural network model provides support for the IR model. Instead of assuming two independent evaluations arising from two distinct processes or systems (i.e., implicit and explicit systems/processes), the network demonstrates that multiple processes underlying evaluation can capture the interaction of earlier processing and more detailed, later processing in the determination of social evaluations. These insights into the mechanisms of human evaluation have significant implications for our understanding of how attitudes are stored, constructed, and changed, affecting how we approach many social and cognitive phenomena.

### Appendix A

#### *Leabra Inhibition, Activation, and Learning Settings*

Leabra has been proposed as a biologically realistic architecture. Leabra's *activation function* results in an S- or sigmoid-shaped pattern of output activation, with minimum and maximum activations. As the level of activation can be thought of as representing the summed firing frequency of a neuron, a node cannot have a negative activation. Thus, possible activations in Leabra range from 0 to 1.

A fundamental aspect of the Leabra architecture is a general mechanism for *inhibition* of the activation of nodes in the network. It is implemented using a version of the k-winners-take-all (kWTA) algorithm (Majani, Erlarson, & Abu-Mostafa, 1989). kWTA inhibition is a method for capturing the impact of inhibition among all the nodes within a layer and calculating how many nodes should be active, given the degree of activation of all the nodes in the layer. One version of the kWTA algorithm is relatively *strict*, allowing no more than k nodes out of a total of *n* (in a layer) to become active at any given time. Other versions of the algorithm are more *lenient* allowing, on *average*, k nodes to be active. In contrast

to the *stricter* version, the more *lenient* version of the kWTA algorithm allows more than  $k$  nodes to be active, if the input activations are sufficiently strong. In Leabra, the strength of inhibition is set for each individual layer.

Inhibition for the hidden layers was based on judgments and experience about the rough number of nodes that needed to be active to learn the desired associations and representations. However, given how inhibition works in Leabra, a fairly wide range of inhibition would work for the hidden layers. Inhibition for the other layers was set to control the number of nodes that should be active for each set of concepts. So we only wanted one profession to be active and multiple attributes to be active, and so we set inhibition in the profession layer so that only one node could be active and we set inhibition in the attribute layer so that as many as three or four attributes could be active.

*Learning* in Leabra combines two different forms of learning: an associative, Hebbian form of learning that captures the correlational or statistical structure of the inputs, and an error-correcting form that enables the network to capture specific task structure (whether an output is correct). Error-correcting learning in Leabra is similar to the better-known delta rule (Woodrow & Hoff, 1960) and enables learning in multilayer networks with hidden units.

Because correlational structure and task structure frequently provide different kinds of information, combining the two kinds of information provides a more powerful learning mechanism (O'Reilly & Munakata, 2000). They are combined by taking a weighted average of the weight change calculated by each learning rule, with the weight given to the associative learning component being much smaller (around .05 or less) than the weight given to the error-correcting component (.95 or higher).

For each of the higher-order semantic knowledge layers, we set a level of inhibition within the layer that would control how many nodes would remain active after processing. The context layer utilized kWTA KV2K inhibition

and  $k$  was set to 2, which drove the layer to prefer to activate only two or so nodes at a time. The profession layer utilized default kWTA inhibition and  $k$  was set to 1, so that only one node would tend to be activated. The attribute layer utilized kWTA KV2K inhibition and  $k$  was set to 3. Hidden Layer 1 utilized kWTA KV2K inhibition and the  $k$  percentage was set to 20%, which drove the layer to activate around 20% of the nodes at one time. Hidden Layer 2 utilized kWTA average inhibition, and the  $k$  percentage was set to 25%.

## Appendix B

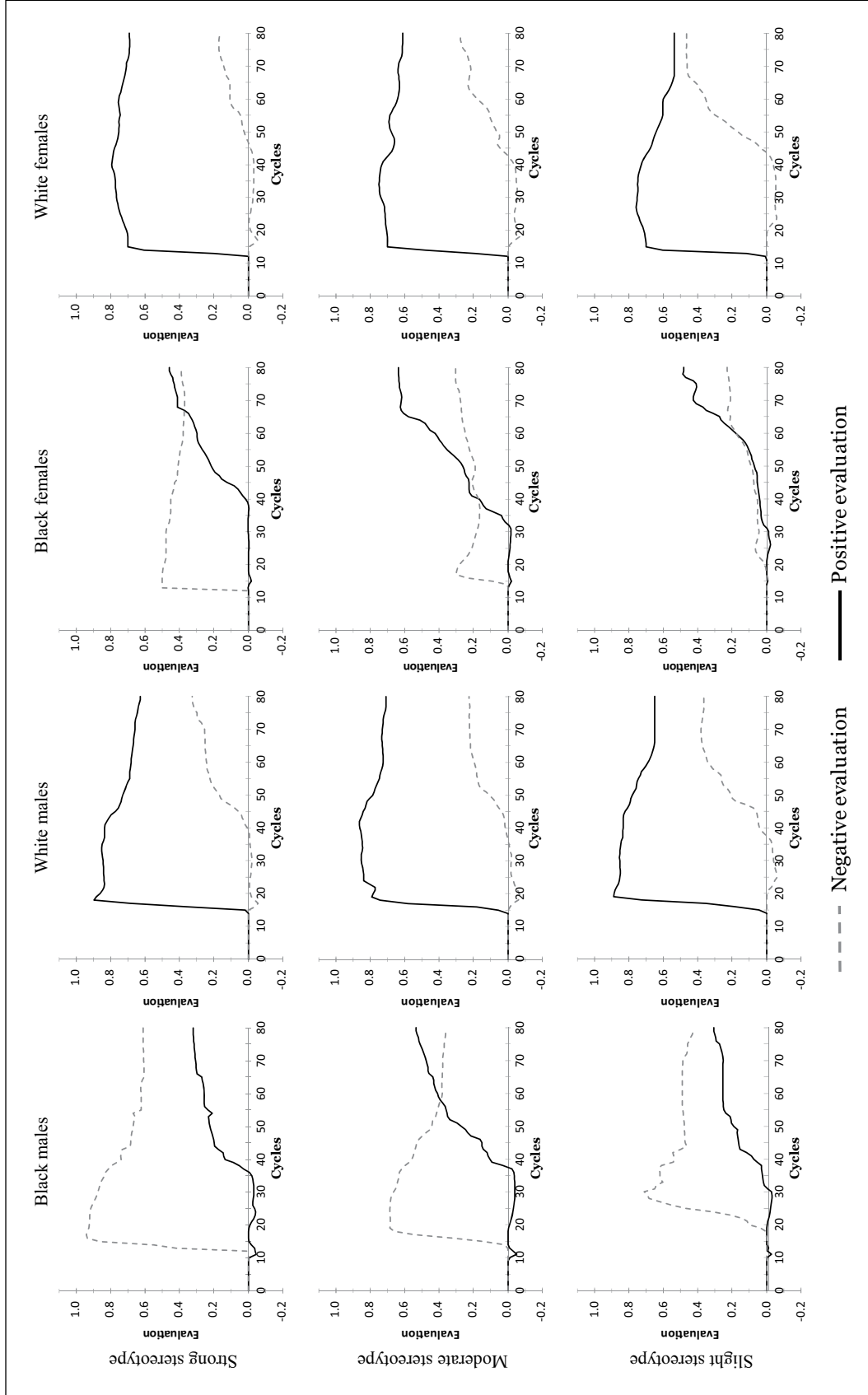
### Learning Details

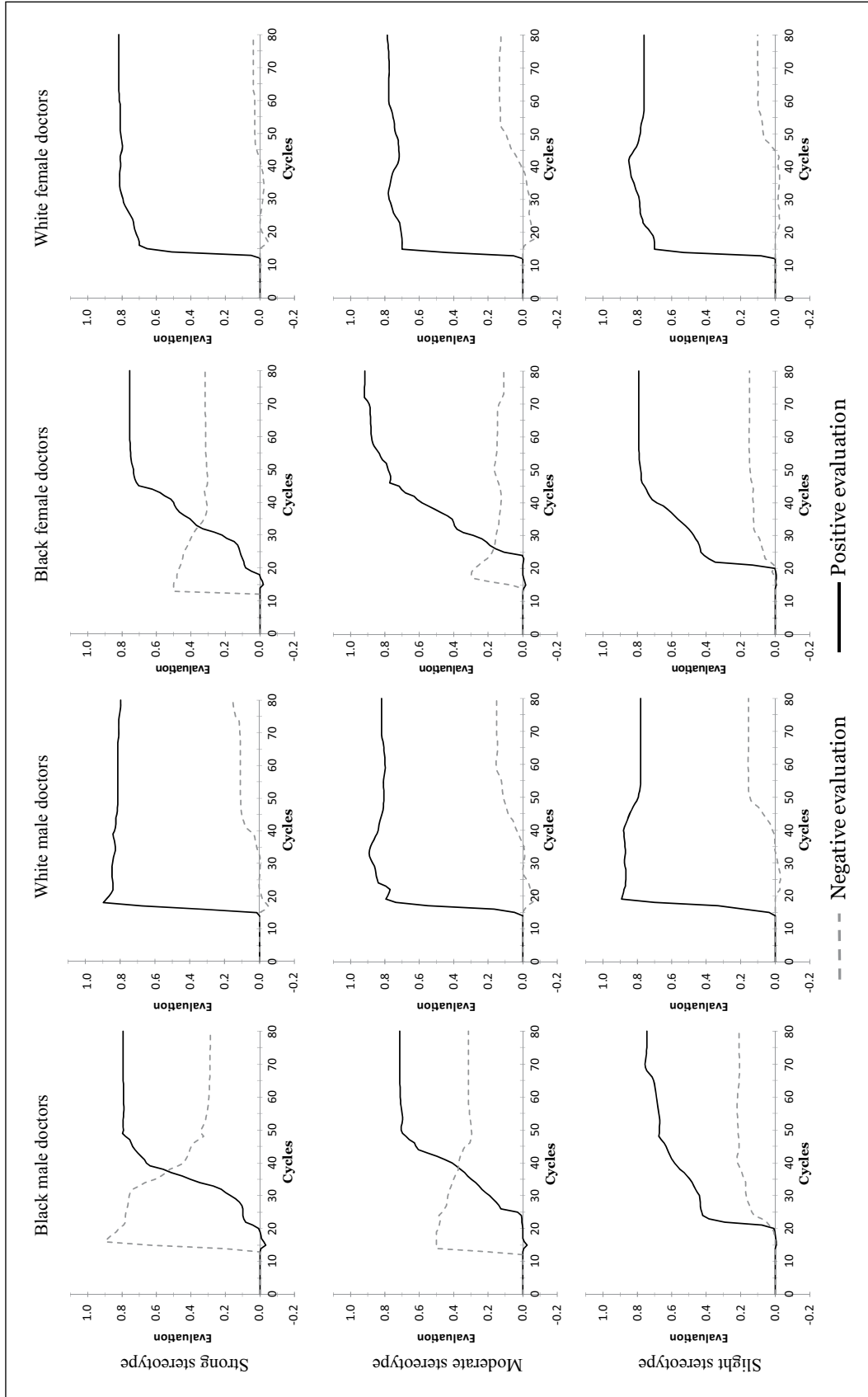
One set of learning parameters was applied to the weights from inputs to race and gender, the weights to race and gender conjunctions, and the weights from race and gender conjunctions to the evaluation layers. These learning parameters were varied between the early and later learning phases. In the early phase of learning, the learning rate was .05 (allowing the network to learn stereotypic racial and gender evaluations), and in the later phase of learning, it was .001 (to preserve the learned stereotypic racial and gender evaluations). The proportion of Hebbian learning was .05, and the proportion of error-correcting learning was .95. A second set of learning parameters applied to all other weights in the network and were constant across learning phases. For these connections, the learning rate was .01, the proportion of Hebbian learning was .001, and the proportion of error-correcting learning was .999.

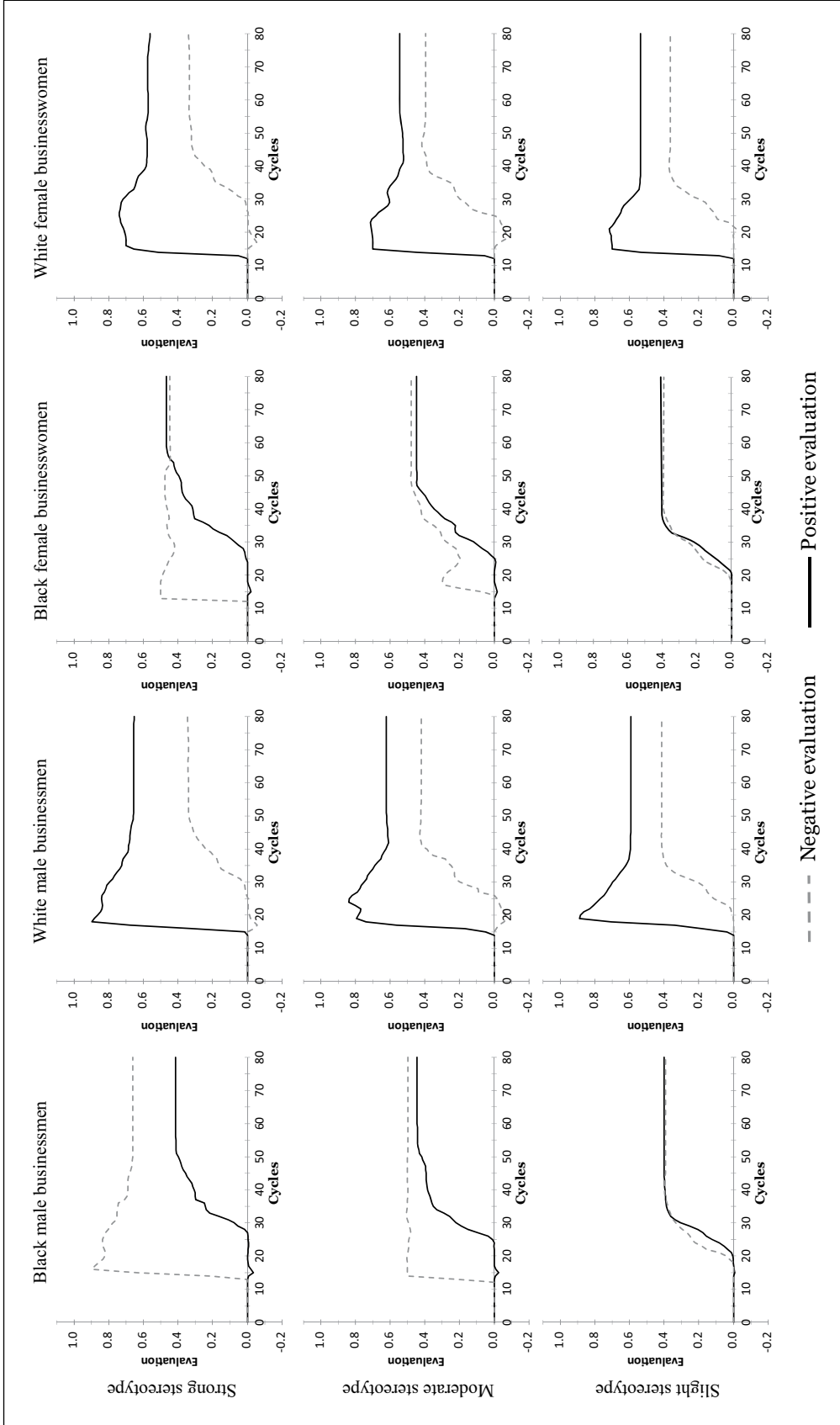
For the connections among unlesioned layers during early training, the learning rate was .05. For late training, the learning rate for all connections was set to .01 (except for the connections that were trained in early training, where the learning rate was set to .001 to preserve the learned race and gender stereotype evaluations).

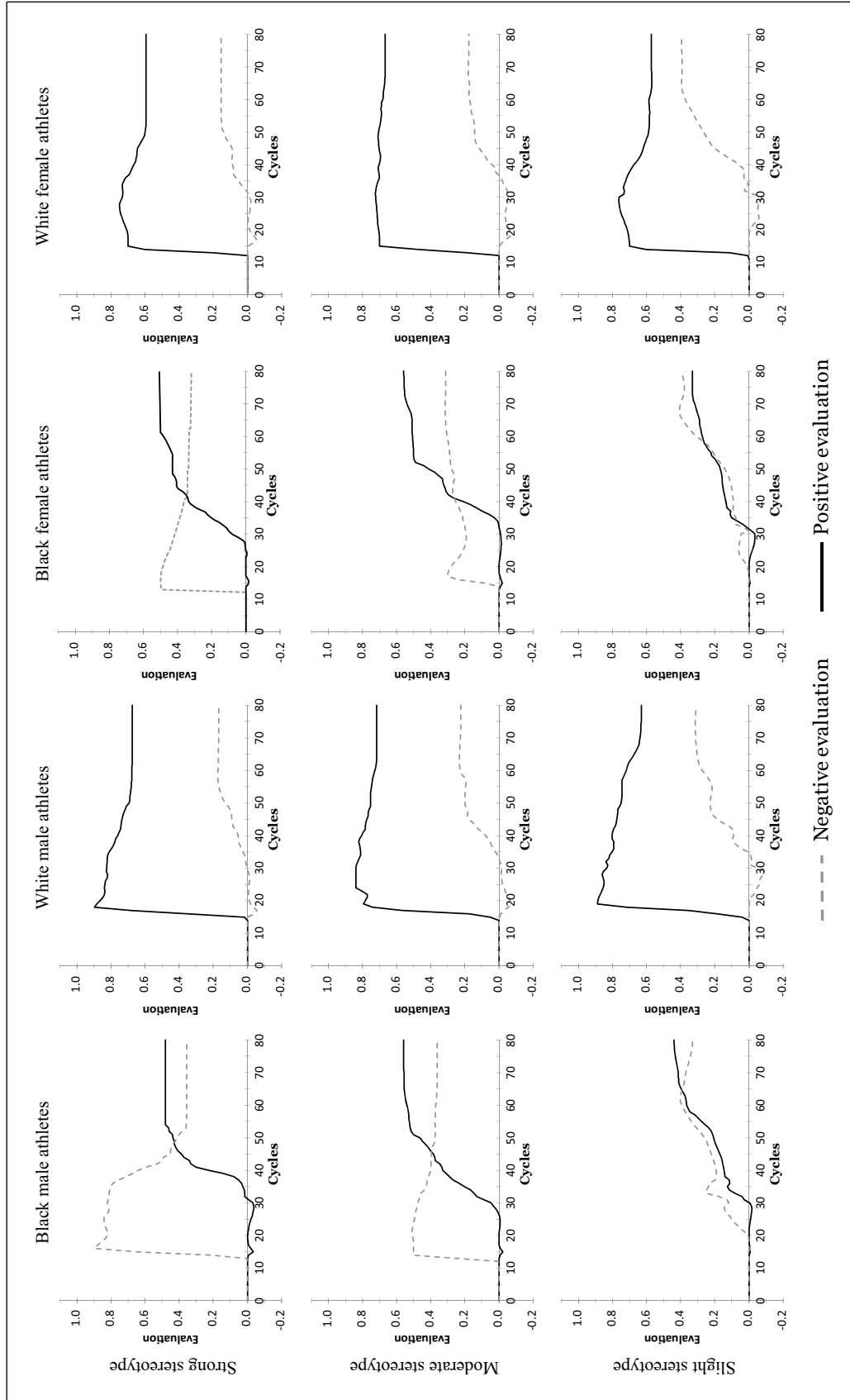
### Appendix C

Evaluation time course graphs by individual and stereotype strength.

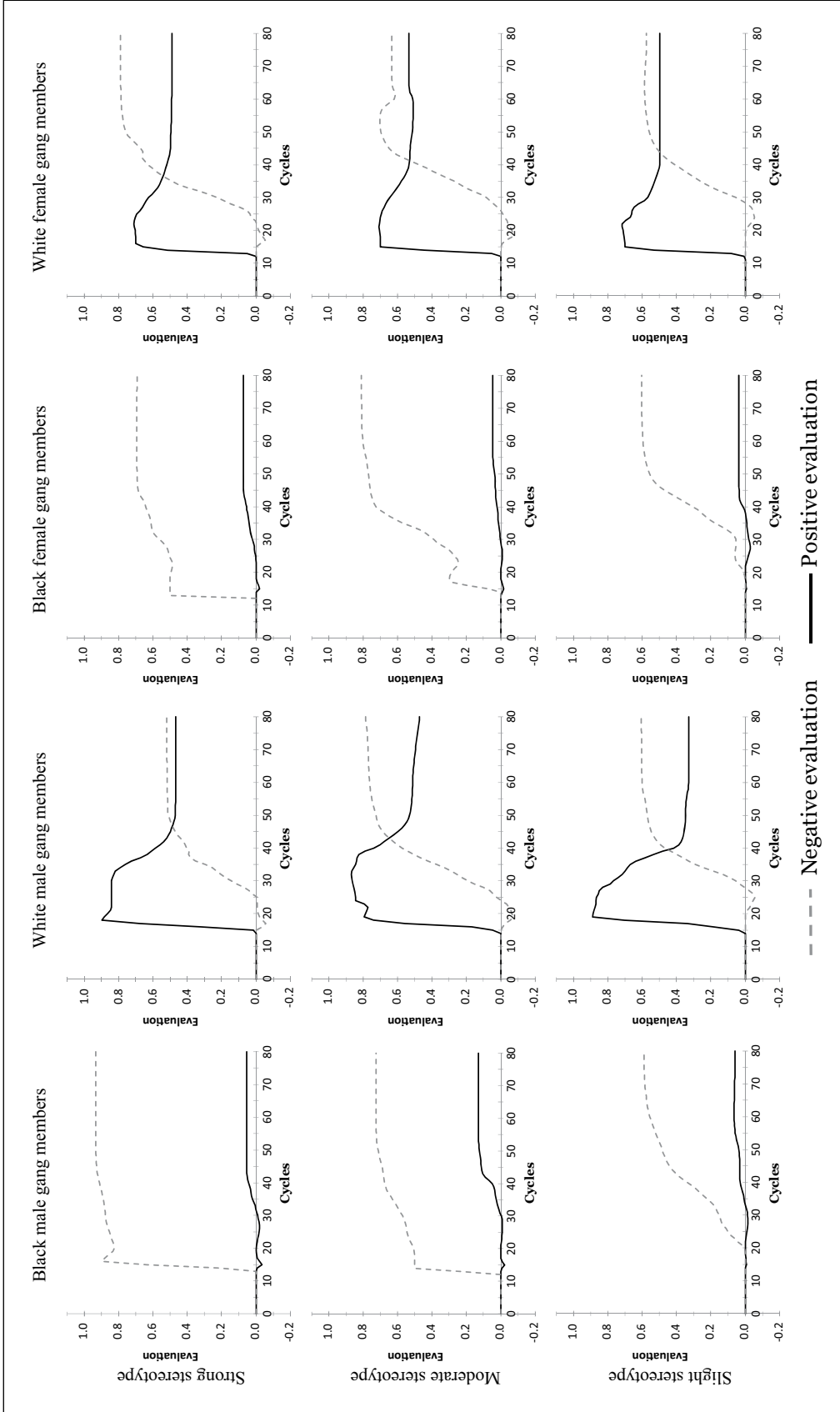






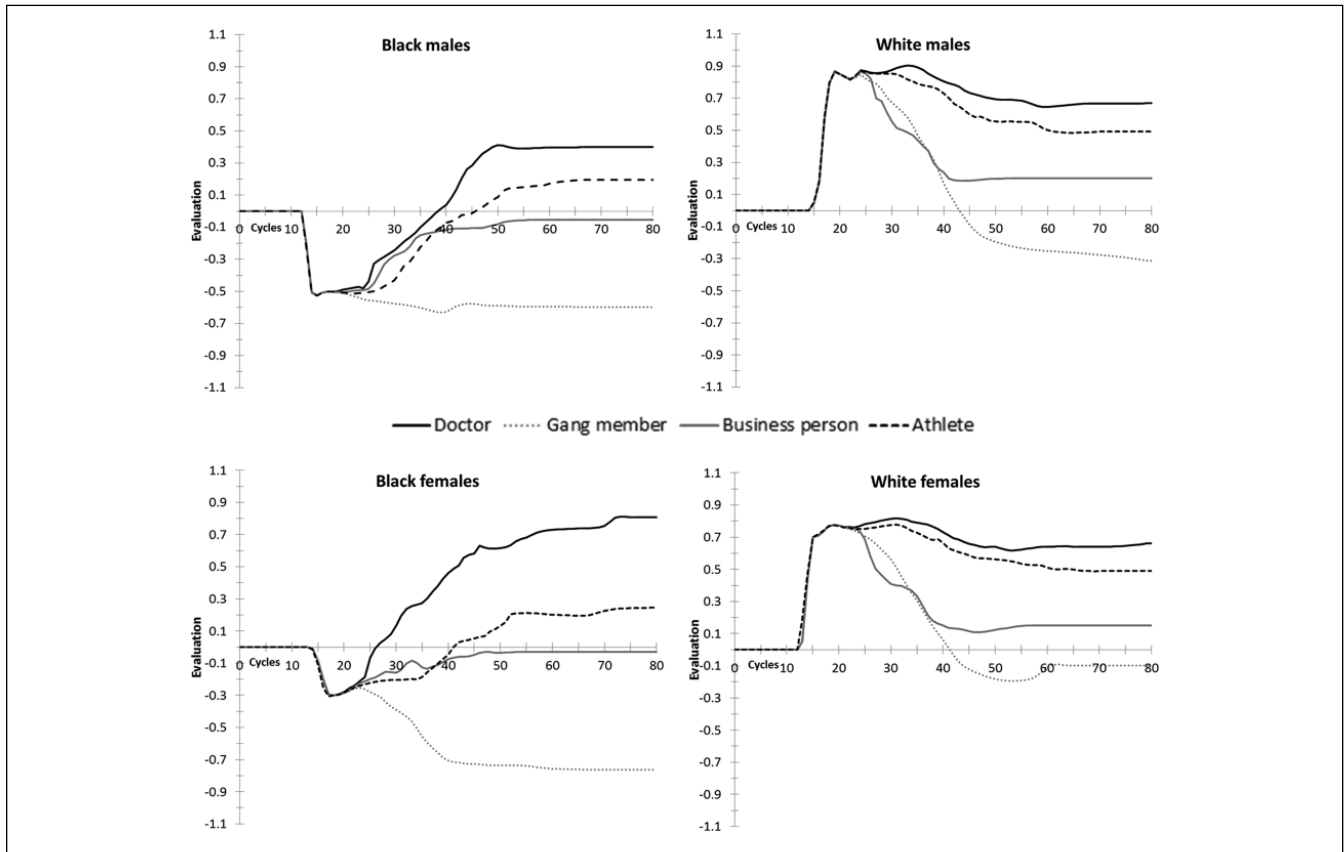






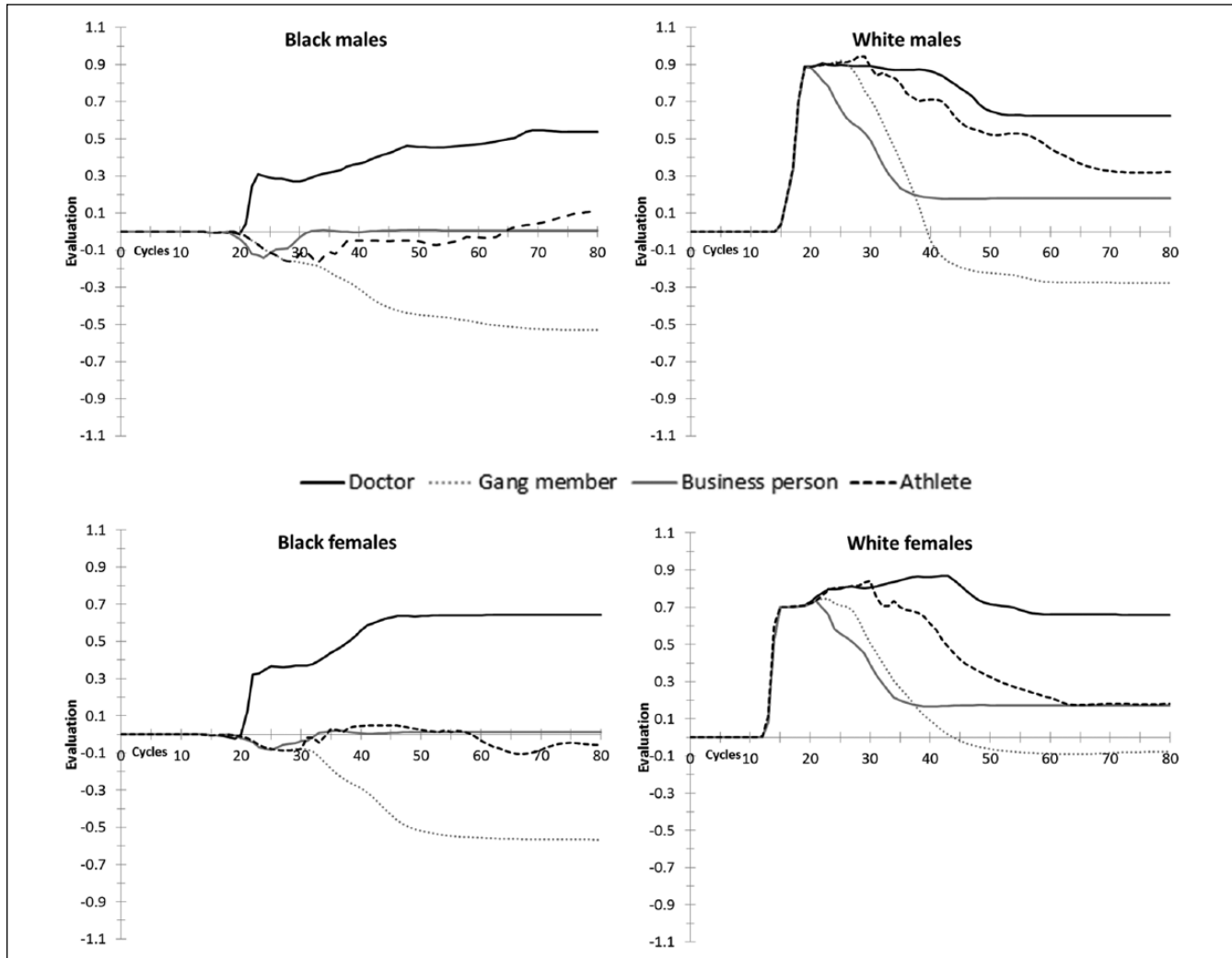
### Appendix D

Net evaluation time courses for moderate stereotypes for all professions, by race and gender.



## Appendix E

Net evaluation time courses for slight stereotypes for all professions, by race and gender.



### Declaration of Conflicting Interests

The author(s) declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### Note

1. We would like to thank two anonymous reviewers for their suggestions that encouraged this additional prediction and simulation.

### References

Aisa, B., Mingus, B., & O'Reilly, R. (2008). The Emergent neural modeling system. *Neural Networks, 21*, 1146-1152.

Amodio, D. M., Bartholow, B. D., & Ito, T. A. (2014). Tracking the dynamics of the social brain: ERP approaches for social cognitive and affective neuroscience. *Social Cognitive and Affective Neuroscience, 9*, 385-393.

Barden, J., Maddux, W. W., Petty, R. E., & Brewer, M. B. (2004). Contextual moderation of racial bias: The impact of social roles on controlled and automatically activated attitudes. *Journal of Personality and Social Psychology, 87*, 5-22.

Bassili, J. N., & Brown, R. D. (2005). Implicit and explicit attitudes: Research, challenges, and theory. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 543-574). Mahwah, NJ: Lawrence Erlbaum.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*, 242-260.

Brewer, M. B., & Feinstein, A. S. (1999). Dual processes in the cognitive representation of persons and social categories. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 254-288). New York, NY: Guilford Press.

- Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, *115*, 401-423.
- Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, *1*, 3-25.
- Conrey, F. R., & Smith, E. R. (2007). Attitude representation: Attitudes as patterns in a distributed, connectionist representational system. *Social Cognition*, *25*, 718-735.
- Correll, J., Wittenbrink, B., Park, B., Judd, C. M., & Goyle, A. (2011). Dangerous enough: Moderating racial bias with contextual threat cues. *Journal of Experimental Social Psychology*, *47*, 184-189.
- Cunningham, W. A., Van Bavel, J. J., Arbuckle, N. L., Packer, D. J., & Waggoner, A. S. (2012). Rapid social perception is flexible: Approach and avoidance motivational states shape P100 responses to other-race faces. *Frontiers in Human Neuroscience*, *6*, 1-7.
- Cunningham, W. A., & Zelazo, P. (2007). Attitudes and evaluations: A social cognitive neuroscience perspective. *Trends in Cognitive Sciences*, *11*, 97-104.
- Cunningham, W. A., Zelazo, P., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition*, *25*, 736-760.
- De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*, 347-368.
- Evans, J. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*, 454-459.
- Evans, J. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255-278.
- Fazio, R. H. (1995). Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 247-282). Hillsdale, NJ: Lawrence Erlbaum.
- Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, *25*, 603-637.
- Fiske, S. T., Lin, M., & Neuberg, S. L. (1999). The continuum model. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 231-254). New York, NY: Guilford Press.
- Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*, 226-241.
- Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*, 247-281.
- Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, Article 59.
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*, 692-731.
- Gawronski, B., & Creighton, L. A. (2013). Dual-process theories. In D. E. Carlston (Ed.), *The Oxford handbook of social cognition* (pp. 282-312). New York, NY: Oxford University Press.
- Gawronski, B., & Sritharan, R. (2010). Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures. In B. Gawronski & B. K. Payne (Eds.), *The handbook of implicit social cognition: Measurement, theory, and applications* (pp. 216-240). New York, NY: Guilford Press.
- 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. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464-1480.
- Huster, R. J., Debener, S., Eichele, T., & Herrmann, C. S. (2012). Methods for simultaneous EEG-fMRI: An introductory review. *The Journal of Neuroscience*, *32*, 6053-6060.
- Ito, T. A., & Urland, G. R. (2003). Race and gender on the brain: Electrocortical measures of attention to race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology*, *85*, 616-626.
- Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, *4*, 533-550.
- Knutson, K. M., Mah, L., Manly, C. F., & Grafman, J. (2007). Neural correlates of automatic beliefs about gender and race. *Human Brain Mapping*, *28*, 915-930.
- Kruglanski, A. W., & Dechesne, M. (2006). Are associative and propositional processes qualitatively distinct? Comment on Gawronski and Bodenhausen (2006). *Psychological Bulletin*, *132*, 736-739.
- Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*, 97-109.
- Kruglanski, A. W., & Orehek, E. (2007). Partitioning the domain of social inference: Dual mode and systems models and their alternatives. *Annual Review of Psychology*, *58*, 291-316.
- Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*, 284-308.
- Majani, E., Erlarson, R., & Abu-Mostafa, Y. (1989). The induction of multiscale temporal structure. In D. S. Touretzky (Ed.), *Advances in neural information processing systems* (pp. 634-642). San Mateo, CA: Morgan Kaufmann.
- Monroe, B. M., & Read, S. J. (2008). A general connectionist model of attitude structure and change: The ACS (Attitudes as Constraint Satisfaction) model. *Psychological Review*, *115*, 733-759.
- O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.
- Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, *12*, 729-738.
- Read, S. J., & Monroe, B. M. (2009). Using connectionist networks to understand neurobiological processes in social and personality psychology. In E. Harmon-Jones & J. S. Beer (Eds.), *Methods in social neuroscience* (pp. 259-294). New York, NY: Guilford Press.

- Sekaquaptewa, D., Vargas, P., & von Hippel, W. (2010). A practical guide to paper-and-pencil implicit measures of attitude. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition* (pp. 140-151). New York, NY: Guilford Press.
- Sinclair, L., & Kunda, Z. (1999). Reactions to a Black professional: Motivated inhibition and activation of conflicting stereotypes. *Journal of Personality and Social Psychology, 77*, 885-904.
- Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review, 11*, 1-18.
- Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*, 220-247.
- Van Bavel, J. J., Xiao, Y. J., & Cunningham, W. A. (2012). Evaluation is a dynamic process: Moving beyond dual system models. *Social & Personality Psychology Compass, 6*, 438-454.
- Van Bavel, J. J., Xiao, Y. J., & Hackel, L. M. (2013). Social identity shapes social perception and evaluation: Using neuroimaging to look inside the social brain. In B. Derks, D. Scheepers, & N. Ellemers (Eds.), *The neuroscience of prejudice and in intergroup relations* (pp. 110-129). Hove, UK: Psychology Press.
- Wegener, D. T., & Petty, R. E. (1997). The flexible correction model: The role of naïve theories of bias in bias correction. In M. P. Sanna (Ed.), *Advances in experimental social psychology* (Vol. 29, pp. 141-208). San Diego, CA: Academic Press.
- Wilson, T. D., Samuel, L., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*, 101-126.
- Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81*, 815-827.
- Wojnowicz, M. T., Ferguson, M. J., Dale, R., & Spivey, M. J. (2009). The self-organization of explicit attitudes. *Psychological Science, 20*, 1428-1435.
- Woodrow, B., & Hoff, M. (1960). Adaptive switching circuits. *Western Electronic Show and Convention, 4*, 96-104.