A Recurrent Connectionist Model of Group Biases

Dirk Van Rooy, Frank Van Overwalle, Tim Vanhoomissen
Vrije Universiteit Brussel, Belgium

Christophe Labiouse
Belgian NFSR Research Fellow & University of Liège, Belgium

Robert French
University of Liège, Belgium

Running Head: Recurrent model of Group Biases

A Recurrent Connectionist Model of Group Biases

Abstract

Major biases and stereotypes in group judgments are reviewed and modeled from a recurrent connectionist perspective. These biases are in the areas of group impression formation (illusory correlation), group differentiation (accentuation), stereotype change (dispersed versus concentrated distribution of inconsistent information), and group homogeneity. All these phenomena are illustrated with well-known experiments, and simulated with an auto-associative network architecture with linear activation update and delta learning algorithm for adjusting the connection weights. All the biases were successfully reproduced in the simulations. The discussion centers on how the particular simulation specifications compare to other models of group biases and how they may be used to develop novel hypotheses for testing the connectionist modeling approach and, more generally, for improving theorizing in the field of social biases and stereotype change.

Petite, attractive, intelligent, WSF, 30, fond of music, theatre, books, travel, seeks warm, affectionate, fun-loving man to share life's pleasures with view to lasting relationship. Send photograph. Please no biochemists.
(Personal ad, New York Review of books, cited in Barrow, 1992, p.2)

The ability to learn about groups and their characteristics is crucial to the way people make sense of their social world. Nevertheless, quite a number of studies have indicated that people can have great trouble learning associations between groups and their attributes and often perceive associations that do not exists. It is generally assumed that these shortcomings or biases are partly responsible for group stereotyping and minority discrimination. Among the most prominent of these group biases are illusory correlation — the perception of a correlation between a group and some characteristics that do not exist (Hamilton & Gifford, 1976; Hamilton & Rose, 1980), accentuation — making a distinction between groups beyond actual differences (Tajfel & Wilkes, 1963; Eiser, 1971), subtyping — the rejection of stereotype-inconsistent information concentrated in a few group members (Hewstone, 1994), and outgroup homogeneity — the perception of outgroups as more homogeneous and stereotypical than the ingroup (Linville, Fisher & Salovey, 1989; Messick & Mackie, 1989).

It is thus of crucial importance to psychologists to understand how these biases are created and how they can be eliminated (Hewstone, 1994). However, many empirical reports on the occurrence of group biases were explained by appeals to what often appear to be rather ad-hoc hypotheses and assumptions. Moreover, the field of group perception has developed largely independent from other important areas in cognition at large and social cognition in particular, including domains such as person perception, impression formation, attribution and attitudes (Hamilton & Sherman, 1996). There have been some recent attempts, however, to provide a common theory of group judgments and shortcomings under the heading of exemplar-based models (Smith, 1991, Fiedler, 1996) or a tensor-product connectionist network (Kashima, Woolcock & Kashima, 2000). The goal of the present paper is to build further on these initial proposals and to present a connectionist model that potentially can explain a wider range of group biases than these earlier attempts. Moreover, the proposed model has already been fruitfully applied to other areas in memory and cognition (for a classic example, see McClelland & Rumelhart, 1986, p. 170), including the domain of social cognition (Read & Montoya, 1999; Smith & DeCoster, 1998; Van Overwalle & Jordens, 2002; Van Overwalle & Labiouse, 2002; Van Overwalle, Siebler & Labiouse, 2002), where it has been applied to encompass and integrate earlier algebraic models of impression formation (Anderson, 1981), causal attribution (Cheng & Novick, 1992) and attitude formation (Ajzen, 1991).

Our basic claim is that a connectionist account of group biases does not require special processing of information as many theories in social cognition posit (e.g., Hamilton & Gifford, 1976; Hastie, 1980). Rather, general information processing characteristics captured in general-purpose connectionist models lead to these biases. What are the characteristics that accomplish this?

First, connectionist models exhibit emergent properties such as the ability to extract prototypes from a number of exemplars (prototype extraction), to recognize exemplars based on the observation of incomplete features (pattern completion), to generalize knowledge about features to similar exemplars (generalization), to adjust to multiple constraints from the external environment (constraint satisfaction), and to lose stored knowledge only partially after damage (graceful degradation). All of these properties have been extensively reviewed in Smith (1996) and McLeod, Plunkett & Rolls (1998). It is clear that these characteristics are potentially useful for any account of group stereotyping. In addition, connectionist models assume that the

development of internal representations and the processing of these representations are done in parallel by simple and highly interconnected units, contrary to traditional models where the processing is inherently sequential. As a result, these systems have no need for a central executive, which eliminates the requirement of previous theories of explicit (central) processing of relevant information. Consequently, biases in information processes are, in principle, due to implicit and automatic mechanisms without explicit conscious reasoning. Of course, this does not preclude people's being aware of the outcome of these preconscious processes.

Second, connectionist networks are not fixed models but are able to learn over time, usually by means of a simple learning algorithm that progressively modifies the strength of the connections between the units making up the network. The fact that most traditional models in psychology are incapable of learning is a significant restriction. Interestingly, the ability to learn incrementally can put connectionist models in agreement with developmental and evolutionary pressures. This implies that group biases emerge from general processes that are otherwise quite adaptive.

Third, connectionist networks have a degree of neurologically plausibility that is generally absent in previous approaches to integration and storage of group information (Anderson, 1981; Ajzen, 1991). While it is true that connectionist models are highly simplified versions of real neurological circuitry and processing, it is commonly assumed that they reveal a number of emergent processing properties that real human brains also exhibit. One of these emergent properties is the integration of long-term memory (i.e., connection weights), short-term memory (i.e., internal activation) and outside information (i.e., external activation). There is no clear separation between memory and processing as there is in traditional models. Even if biological constraints are not strictly adhered to in connectionist models of group prejudice, interest in the biological implementation of social cognitive mechanisms has indeed started to emerge (Adolphs & Damasio, 2001; Ito & Cacioppo, 2001; Ochsner & Lieberman, 2001) and parallel the increasing attention paid to neurophysiological determinants of social behavior.

This article is organized as follows: First, we will describe the proposed connectionist model in some detail, giving the precise architecture, the general learning algorithm and the specific details of how the model processes information. In addition, a number of other less well-known emergent properties of this type of network will be discussed. We will then present a series of simulations, using the same network architecture applied to a number of important biases in group judgments, including illusory correlation, accentuation, stereotype change and homogeneity. Our review of empirical phenomena in the field is not meant to be exhaustive, but is rather designed to illustrate how connectionist principles can be used to shed light on the processes underlying group judgments.

While the emphasis of the present article is on the use of a particular connectionist model to explain a wide variety of group biases, previous applications of connectionist modeling to social psychology (e.g., Smith & DeCoster, 1998; Read & Montoya, 1999; Van Overwalle, 1998; Van Overwalle & Jordens, 2002) are also mentioned and compared to the present approach.. Finally, we will discuss the limitations of the proposed connectionist approach and discuss areas where further theoretical developments are under way or are needed. Ultimately, what we would like to accomplish in this paper is to create a greater awareness that connectionist principles could potentially underlie diverse shortcomings in group judgments, as a natural consequence of the basic processing mechanisms in these adaptive cognitive systems.

A Recurrent Model

Throughout this paper, we will use the same basic network model - namely, the recurrent

auto-associator developed by McClelland and Rumelhart (1986, 1988). This model has already gained some familiarity among psychologists studying person and group impression (Smith & DeCoster, 1998; Queller & Smith, 2002), causal attribution (Read & Montoya, 1999) and many other phenomena in social cognition (Van Overwalle & Labiouse, 2002; Van Overwalle, Siebler, & Labiouse, 2002). We decided to apply a single basic model to emphasize the theoretical similarities that underlie group biases with a great variety of other processes in cognition. In particular, we chose this model because it is capable of reproducing a wider range of phenomena than other connectionist models, such as feedforward networks (see Read & Montoya, 1999), constraint satisfaction models (Kunda & Thagard, 1996; see also Van Overwalle, 1998), or tensor-product models (Kashima, Woolcock & Kashima, 2000).

The auto-associative network can be distinguished from other connectionist models on the basis of its architecture (how information is represented in the model), its learning algorithm (how information is processed in the model) and its testing procedure (how knowledge in the network is retrieved). We will discuss these points in turn.

Architecture

The generic architecture of an auto-associative network is illustrated in Figure 1. Its most salient property is that all nodes are interconnected with all of the other nodes. Thus, all nodes send out and receive activation. The nodes in the network can represent groups, attributes implied in the descriptions of the group, as well as episodic information on specific behaviors and so on. This, in fact, reflects a localist representation where each node represents a single symbolic concept, in contrast to a distributed representation where each concept is represented by a pattern of activation across a set of nodes (Thorpe, 1994). We elaborate on the differences between these two representation schemes in the section on Fit and Model Comparisons.

Information Processing

In a recurrent network, processing information takes place in two phases. During the first activation phase, each node in the network receives activation from external sources. Because the nodes are interconnected, this activation is spread throughout the network in proportion to the weights of the connections to the other nodes. The activation coming from the other nodes is called the internal input (for each node, it is calculated by summing all activations arriving at that node). This activation is further updated during one or more cycles through the network. Together with the external input, this internal input determines the final pattern of activation of the nodes, which reflects the short-term memory of the network. Typically, activations and weights have lower and upper bounds of –1 and +1.

In the linear version of activation spreading in the auto-associator that we use here, the final activation is the linear sum of the external and internal input after a single updating cycle through the network. In non-linear versions used by other researchers (McClelland & Rumelhart, 1986; Smith & DeCoster, 1998; Read & Montoya, 1999), the final activation is determined by a non-linear combination of external and internal inputs updated during a number of internal cycles (for mathematical details, see McClelland & Rumelhart, 1988, pp. 161—169). During our simulations, however, we found that the linear version with a single internal cycle (see p. 167) often reproduced the observed data at least as well. Therefore, we used this linear variant of the auto-associator for all the reported simulations. We will discuss later why the linear variant might have been so efficient.

After the first activation phase, the recurrent model enters the second learning phase in which the short-term activations are consolidated in long-term weight changes of the connections. Basically, these weight changes are driven by the error between the internal input

received from other nodes in the network and the external input received from outside sources. This error is reduced in proportion to the learning rate that determines how fast the network changes its weights (typically between .01 and .20). This error reducing mechanism is known as the *delta algorithm* (McClelland & Rumelhart, 1988, p. 165—166).

For instance, if the external input on group membership is underestimated (e.g., because the internal input predicts a weak or ambiguous member of the group while that person is actually is a very typical member), the connection weights with the group unit are increased to reduce this discrepancy. Conversely, if the external input on group membership is overestimated (e.g., because the internal input predicts an overly idealized prototypical member), the weights are decreased. These weight changes allow the network to better approximate the external input. Thus, the delta algorithm strives to match the internal predictions of the network as closely as possible to the actual state of the external environment, and stores this information in the connection weights.

Testing

To test the knowledge embedded in the connections of the network, we applied a procedure analogous to measuring human responses, that is, where participants are cued with questions on the experimental stimulus material learned previously. To accomplish this, some concepts in the network served as a cue to retrieve related material in the network (e.g., a group label may serve as a cue to estimate group attributes), by turning the activation of the cue on to +1. A series of adjustments by the learning algorithm during learning results in a certain configuration of connection weights in the network. This configuration determines how activation flows through the network and activates related concepts. The degree to which these other, related concepts are activated is taken as a measure of retrieval in memory, and may be indicative of various responses such as estimation (e.g., of groups attributes) or recognition (of group member's behaviors).

A Recurrent Implementation of Group Biases: Illusory Correlation

To provide some background to our specific implementation of group biases, we illustrate its major characteristics with the phenomenon of illusory correlation. Illusory correlation occurs when perceivers erroneously see a relation between categories that are actually independent. For instance, minorities or outgroups are often stereotyped with bad characteristics, although these characteristics sometimes occur in equal proportions in the ingroup. The earliest demonstration of illusory correlation in a group context comes from a study by Hamilton and Gifford (1976). Participants read about members of two groups A and B that engaged in the same ratio of desirable to undesirable behaviors (9:4), but twice as many behaviors referred to members of group A than to members of group B. Although there was no objective correlation between group membership and desirability of behavior, participants showed greater liking for the majority group A than for the minority group B. In sum, the typical finding in illusory correlation research is *decreased evaluation* for minority group B, together with *increased memory* for undesirable group B behavior (for reviews see Hamilton & Sherman, 1989; Mullen & Johnson, 1990).

To account for these two distinct effects in illusory correlation, we introduce a recurrent connectionist model that permits encoding and retrieval of two types of information. One type of information concerns some salient regularity or attribute about the group (such as desirability) and is assumed to underlie the evaluative (i.e., likeability) judgments in illusory correlation. The other type of information involves specific episodic knowledge about the behavioral items and is

assumed to account for the memory effects.

We have chosen a "localist" encoding scheme, that is, each piece of information (or concept) is represented by a single node. Figure 2 shows how the two groups, A and B, are each represented by a *group node* and how the implied attribute (i.e., desirable or undesirable) is represented by two separate *attribute nodes*. Two separate unitary attribute nodes were taken rather than a bipolar attribute node (with positive and negative activation to represent desirable and undesirable stimuli respectively) because our evaluations about groups are not represented as a single point on a one-dimensional construct, but are probably more mixed and complex including both positive and negative instances of the attribute (Wittenbrink, Judd & Park, 2001). In order to explain memory for specific statements presented, we also include *episodic nodes* that reflect the specific (i.e., behavioral) information contained in the statements. Episodic memory refers to information about particular events that have been experienced (Tulving, 1972). The important advantage of episodic nodes is that they preserve information about discrete events in the network. In sum, we assume that the unique meaning of each behavioral statement in an illusory correlation experiment is encoded at two levels: Its evaluative meaning ("the behavior is good") and its unique episodic meaning ("helps an old lady across the street"). By representing different aspects (or features) of each piece of information over two nodes, evaluative and episodic, this model in fact uses a semi-localist encoding scheme.

It is instructive to note that although in principle, in an auto-associative network, all interconnections between all nodes play a role, to understand the present simulations, the reader should focus mainly on the connections between different sets of nodes (e.g., between attribute nodes, episodic nodes, and group nodes) that are of most relevance for explaining group biases, while the lateral interconnections linking the same sets of nodes are less relevant (contrary to spreading activation models of impression formation, e.g., Hastie & Kumar, 1979). The connections between episodic nodes and group nodes (in both directions) are collectively termed *episodic connections*, while the connections between evaluative attribute nodes and the group nodes (in both directions) are termed *evaluative connections*.

The delta learning algorithm gives rise to a number of emergent properties that are used to explain all the effects associated with group biases. Below, we describe two of the most important properties and illustrate their effect on the illusory correlation bias.

Acquisition Property and Sensitivity to Sample Size

According to the delta algorithm, the more an attribute such as an (un)desirable behavior is presented with information on group membership, the stronger the connection between the corresponding (un)desirability attribute node and group node becomes. This illustrates an important property of the delta learning algorithm, namely that as more confirmatory information is received, the connections gradually grow in strength. We call this the *acquisition property*. Thus, in the beginning phases of learning (before asymptote is reached), the connection weights reflect the amount of evidence, that is, the network is sensitive to sample size.

The sensitivity to sample size of the delta algorithm has already been exploited in the earlier associative learning models that preceded connectionism, such as the popular Rescorla-Wagner (1972) model of animal conditioning and human contingency judgments. This model predicts that when a cue (i.e., conditioned stimulus) is followed by an effect (i.e., unconditioned stimulus), the organism integrates this information resulting in a stronger cue-effect association and more vigorous responding when the cue is present. In humans, this also results in stronger judgments of the causal influence of the cue (see Baker et al. 1989; Shanks, 1985a, 1987, 1995; Shanks, Lopez, Darby & Dickinson, 1996; Van Overwalle & Van Rooy, 2001). Sample size effects have also been documented in many areas of cognition. For instance, when receiving

more supportive information, people tend to hold more extreme impressions about other persons (Anderson, 1967, 1981), make more extreme causal judgments (Baker, Berbier & Vallée-Tourangeau, 1989; Shanks, 1985a, 1987, 1995; Shanks et al., 1996), make more polarized group decisions (Fiedler, 1996; Ebbesen & Bowers, 1974), endorse more firmly an hypothesis (Fiedler, Walther & Nickel, 1999), make more extreme predictions (Manis, Dovalina, Avis & Cardoze, 1980) and agree more with persuasive messages (Eagly & Chaiken, 1993).

How does the acquisition property explain illusory correlation? The mechanism is straightforward. Because of the larger sample size in the majority group A, its evaluative connections are stronger at the end of learning than the corresponding connections for the minority group B. Thus, both the connections with desirability and undesirability are stronger for group A than for group B. As a result, the relative proportion of desirable versus undesirable information is more clearly encoded in the evaluative connections of the network for majority group A than for minority group B, resulting in a more favorable impression overall for the majority group A. In addition, this also means that the mental representation of the majority group A, in contrast to the minority group B, will consist of well-established connections between group membership and (un)desirability of behavior, so that the perceiver can form a relatively correct impression of the majority group. It is important to note, however, that when both groups become larger, the relative advantage of the majority group A will be lost as the evaluative connections of both groups will reach their asymptote. However, this is not typical of illusory correlation experiments, where the number of statements is most often less than 20 for each group.

Figure 3 depicts an simulated example of this process. We focus here on the connections from the group nodes to the desirability and undesirability nodes. We simulated the presentation of desirable and undesirable statements on the groups by activating, for each statement, the respective desirability nodes and group nodes. After each statement, we tested the strength of the evaluative connections by cuing each group node and measuring how much of the activation was spread to the desirability nodes. As can be seen, the strength of the evaluative connections increases as a function of the growing number of statements. The top half of the figure shows this for majority group A, the bottom half for minority group B. Every time a statement is presented (for instance "John helps an old lady across the street"), the simulated evaluation increases. Although the increase with each statement is equal for both groups, the larger amount of statements (larger sample size) for group A results in stronger connections and a larger difference between desirable and undesirable evaluations for the majority group A than for the minority group B ($D_a > D_b$ in the figure). As a minor point, note that the evaluations after four trials in Figure 3 differ between groups A (.36) and B (.25) because the lateral connections between the nodes also differ in number between groups (these curves would have been exactly similar if lateral connections were omitted such as in feedforward network models, discussed later).

Competition Property and Discounting

In order to explain enhanced memory for negative behaviors and for minority behaviors, we now turn to the episodic nodes that reflect memory during illusory correlation. We propose that a memory advantage for these infrequent behaviors in recognition measures, where episodic nodes presumably serve as retrieval cue to remember the group, may in part be produced by what has been termed the *competition property* of the delta learning algorithm (Shanks, 1995; Van Overwalle & Van Rooy, 1998). The term "competition" stems from the associative learning literature on animal conditioning and causality judgments where it is also known as blocking (Rescorla & Wagner, 1972; Shanks, 1995), and should not be confused with other usages in the

connectionist literature such as competitive networks (McClelland & Rumelhart, 1988).

The competition property favors features that are more diagnostic than others, which are disfavored. A typical example is discounting in causal attribution. When one cause acquires strong causal weight, perceivers tend to ignore alternative causes. As noted by several researchers (Read & Montoya, 1999; Van Overwalle, 1998), competition in learning is also a hallmark of the Rescorla-Wagner (1972) model. Competition is a robust finding in empirical research on animal conditioning (Kamin, 1969), human causal learning (Shanks, 1985b) and causal attribution (Hansen & Hall, 1985; Kruglanski, Schwartz, Maides, & Hamel, 1978; Van Overwalle & Van Rooy, 1998; Read & Montoya, 1999; Wells & Ronis, 1982).

How does the competition property explain enhanced memory in illusory correlation? The basic mechanism behind competition is that only a limited amount of connection strength is available during learning. In illusory correlation, this limitation is a function of the external activation of a group node (limited to +1 in the present case to reflect group membership) and of the internal activation received from other evaluative and episodic nodes, and affects the connections from the evaluative and episodic nodes to the group nodes (see upward arrows in Figure 2). Because the delta algorithm seeks to match internal with external activations, the internal activation received from the evaluative and episodic nodes (and hence also their connection weights) cannot grow out of bounds, as their sum is limited by the upper value of the external activation of the group node. Stated differently, given an upper external activation of +1 of a group node, the internal activation sent by episodic and evaluative nodes to that group node is limited. To the extent that the sum of this internal activation approaches or exceeds the upper bound, these nodes have to compete for connection weights and the growth of their connections is blocked or reduced. A consequence of this is that strong group→attribute connections contribute much more in approaching or exceeding the upper limit than weaker group→attribute connections, and so tend to discount or block the further growth of the episodic→group connections more.

To take the acquisition example of Figure 3, the connection weight from the desirable node to group A at the end of learning is 0.74, and hence leaves only 0.26 activation available before the upper bound of the external activation (+1) of the group node will be exceeded. Conversely, the same weight for group B is only 0.51, and there is thus much more room (0.49) for increasing the weights of the desirability and other nodes. Thus, because of the stronger attribute→group connections of group A, the episodic→group connections of this group will be much more discounted, resulting in reduced memory for behavioral episodes (see schematic illustration in Figure 4, top panel). In contrast, because of the weaker attribute→group connections of group B, the episodic→group connections of this group will be less discounted, so that they can gain more connection weight resulting in enhanced memory (see Figure 4, bottom panel). By the same mechanism, because the desirable→group connections are larger than the undesirable→group connections, the episodic→group connections of positive behaviors will be weaker than those of the negative behaviors, resulting in an increased memory for negative behavior. In sum, the competition property generates a memory advantage, not for paired distinctive stimuli like the distinctiveness account would predict, but separately for undesirable and minority behaviors because of their infrequency [1].

Summary. The acquisition and competition properties of the delta learning algorithm shape the connections between group nodes, attribute nodes and episodic nodes as information is provided about the groups. Essentially, these properties describe different ways in which a growing number of observations affect connections in the network. The acquisition property describes how the attribute connections grow stronger as a function of a growing sample size,

and so enables a preponderance of desirable behaviors in the majority group much quicker than in the minority group. As a consequence, no paired distinctive stimuli are necessary to produce the illusion correlation effect. This implication has received empirical support from recent studies (Shavitt, Sanbonmatsu, Smittipatana & Posavac, 1999; Van Rooy, 2001), that have showed that the effect is obtained without any negative behavioral information on the minority group, contrary to the distinctiveness hypothesis. The competition property describes how stronger attribute→group connections inhibit the development of weaker episodic→group connections. This latter property plays a role in the memory advantage for infrequent behaviors.

Other Relevant Theories

Before applying our recurrent implementation to several group biases of interest, we will first briefly compare the recurrent approach with the two most relevant models that have been proposed in the recent past to explain group biases such as the illusory correlation effect.

Exemplar Models

Perhaps the most well known theoretical approach to explain group biases was inspired by recent exemplar models of memory (Fiedler, 1996; Nosofsky, 1986; Smith, 1991; Smith & Zárate, 1992). According to exemplar models, perceivers store single exemplars of behaviors in memory. To make a judgment about a target stimulus (e.g., a group), perceivers form a composite estimate of activated memory traces of the stored exemplars that are highly similar to the target stimulus. Thus, group judgments are based on specific exemplars that are retrieved from memory and aggregated. In the exemplar models of Smith (1991) and Fiedler (1996; Fiedler, Kemmelmeier & Freytag, 1999) that provide the most detailed accounts of social judgments, this aggregation is based on a simple or weighted linear summation. Such an aggregation process will cancel out unsystematic perceptual or encoding errors between the exemplars, and will reinforce systematic variance. An important consequence is that less error variance is left in the aggregate, the larger the amount of observations. This is important, because as less error variance is left, then perceptions of the group become more accurate, alleviating the tendency to make biased judgments. Hence, exemplar theories essentially explain many group biases by information loss or insufficient evidence, and predict that increasing the encoding of actual group information can alleviate judgmental shortcomings. Like the present recurrent network, they are thus sensitive to sample size differences.

One major difference with our recurrent approach is that in exemplar models, information about behavioral episodes and their trait or evaluative implication is solely encoded at the exemplar level, while (aggregated) attributes are computed at retrieval. In addition, because the evaluative attributes are computed from the exemplars, it is predicted that there should be a positive correlation between judgment and memory, that is, lower liking for minority group B should result in lower recall for the behavior exemplars also (Fiedler, Russer & Gramm, 1993). This stands in contrast to illusory correlation research that shows increased memory for group B exemplars (for recent evidence, see Hamilton, Dugan & Trollier, 1985; McConnell, Sherman & Hamilton, 1994; Stroessner, Hamilton & Mackie, 1992). This observed discrepancy between judgment (decrease) and memory (increase) was overcome in the implementation of our recurrent network by encoding both types of exemplar and attribute information, and by the competition property of the delta algorithm. Such competition mechanism does not exist for exemplar-based models. Another difference is that our model does not require random noise in the encoding of the information to explain group biases, because the delta algorithm is an acquisition device that in itself is sensitive to sample size differences.

Tensor-Product Model

Kashima, Woolcock and Kashima (2000) proposed a connectionist model of group impression formation and change that they called the tensor-product model. It encodes different aspects of social information, including the person, the group to whom he or she belongs to, as well as the specific action or characteristics they express. Like our model, it assumes that this information is encoded in memory as connections between sets of nodes reflecting these different aspects. Thus, aggregation of episodic information takes place during encoding rather than during retrieval, through the strengthening of the connections between nodes. The Hebbian learning algorithm, which involves a weighted linear summation of information, determines weight adjustments.

One of the most important differences with our recurrent model is that the tensor-product model does not say anything about recall of specific behavioral information. In principle, all episodic information is immediately aggregated in the connections and then lost after activation fades away. Moreover, all connections between two nodes are symmetric; while they can differ in the recurrent model depending on the direction in which the activation is spread between the nodes.

Another difference is that the Hebbian algorithm applied in the tensor-product model is not bounded or normalized as it simply keeps on accumulating the weights from previous learning, forcing them beyond -1 and +1. Normalizing takes place only during judgment, for instance, by retrieving appropriate low-end and high-end anchors to calibrate the current judgment. Although research has revealed that people shift their standards of judgment as they think of members of different social groups (e.g., an assertive person is judged "very assertive" as a women but only "mildly assertive" as a man; Biernat & Manis, 1994), this does not necessarily imply that anchors are used during retrieval only. Perhaps, anchors are also used during encoding. For instance, group stereotypes and norms may act as a context against which novel information about members is assessed. This latter anchoring process is outside the scope of the model (although the delta algorithm can address this through the competition property, for more details see Van Overwalle & Van Rooy, 1998). Perhaps, the most important limitation of non-normalized learning in the tensor-product model is that it does not allow limiting activation at each learning trial, so that competition cannot take place. As a consequence, the discrepancy between information loss and increased memory cannot be accounted for.

Overview of the Simulations

In the next sections, we will describe a connectionist simulation of several biases in group judgments. An overview of these simulations is given in Table 1, together with the major connectionist property that drives the bias.

Model Parameters

For all simulations, we used the linear auto-associative recurrent network described above, with parameters for decay and excitation (for internal and external input) all set to 1, and with one internal activation cycle. Node activation was determined by the linear sum of all internal and external inputs received at a node (McClelland & Rumelhart, 1988, p. 167). This effectively means that node activation was solely determined by the sum of external and internal activation received (after one internal cycle through the network) and that activation decay does not play any role in the simulations, nor does multiple cycles of activation updating. These characteristics are identical to recent work by Van Overwalle and colleagues (Van Overwalle & Labiouse, 2002; Van Overwalle, Siebler & Labiouse, 2002).

As noted earlier, the external input was typically bound between -1 and +1, and this leads (with some small deviations) to a similar limitation of the connection weights. Such a limit is not only important to produce competition, but it is also instrumental in preventing the connections growing without bound. The learning rate that determines the speed by which the weights of the connections are allowed to change was set to 0.15. In order to generalize across a range of presentation orders, each network was run for 50 different random orders, thus simulating 50 different "participants". All connection weights were initialized at starting values of zero. Unless otherwise stated, most of the effects are relatively robust to changes in parameters and stimulus distributions. Consequently, for simplicity of presentation, a smaller number of trials were used in some of the simulations than is typically the case in actual experiments.

Testing the Network

To test the judgments, categorizations and memory arising from the network, we measured how much some concepts in memory are able to activate other concepts. Thus, we simply activated node x and looked at the resulting internal activation of node y. For instance, to measure the attributes associated with a group, the group node that serves as cue is primed by turning on its external activation to 1. This activation then spreads to related nodes in proportion to the weights of their connection, and the resulting internal activation (or output activation) is then measured (i.e., read off) from the attribute nodes. This resulting activation can acquire any value between approximately -1 and +1, depending on the weight and direction of the connections. In addition, for some judgments, the activation of some nodes was subtracted (i.e., got a negative sign) in the calculation of the overall assessment (e.g., the activation of opposing valences was subtracted from each other to obtain an overall evaluation measure). No external input activation was provided to the "measurement" nodes, because zero activation is considered a neutral resting activating level just in the middle of the -1 and +1 bounds for activation. (Providing any extra external activation to nodes that serve as measure would bias the response in a given positive or negative direction, and that is of course undesirable).

To simplify, one might think of this procedure as testing the strength of the connection between nodes x and y, because the lateral connections between the same types of nodes quite often (but not always) play only a minor role. (In the section on model comparisons discussed later, we will demonstrate that feedforward networks without such lateral connections often do as well as recurrent networks). Note that if more than one output activation was read off, we averaged the results so that the total output activation remained between the -1 and +1 bounds.

We used the same basic cue and measurement nodes throughout all the simulations. Unless stated otherwise, for central tendency measures of the group (e.g., liking, frequency estimates), the group nodes were turned on and the differential output activation of the attribute nodes was read off. For instance, the resulting activation of the undesirable attribute was subtracted from the resulting activation of the desirable attribute to obtain an overall likeability estimate. For central tendency measures of exemplars (e.g., attitude position of statements, typicality of members), we used exemplar nodes as cues instead of group nodes. Recognition in the assignment task was simulated by first activating each episodic node, and reading off the resulting activation of the group node. Finally, for measures of variance, the same cues and measurement nodes were used as for the central tendency measure of the group, but the resulting activation of the two opposing attribute nodes was summed instead of subtracted. For more details, we refer to the each of the simulations and associated tables (were measurement nodes are denoted by ?).

All the results of the simulations are presented together with observed means from an

illustrative experiment. Like many authors in the associative learning domain (e.g., Nosofsky, Kruschke & McKinley, 1992; Shanks, 1991), we assume that the relationship between the activation resulting from such a test and the judgments by participants is monotonic.   Hence, given that we are mainly interested in patterns of the simulated values, the simulated means per condition are estimated to fit as closely to the human data using linear regression (i.e., we linearly regressed all simulation means onto all human means and use that regression to compute human-like values for the simulation). This procedure also enables us to demonstrate visually the fit of the simulations.

Group Impressions

How do perceivers develop a stereotyped impression of a group? Of the many processes that may contribute to a biased group perception, we focus on illusory correlation as a consequence of sample size differences (as introduced earlier) and as a consequence of prior expectancies (to be discussed later).

Simulation 1: Size-based Illusory Correlation

As noted earlier, size-based illusory correlation refers to the tendency to perceive minority groups as more negative than majority groups, despite an equal preponderance of desirable behaviors in the two groups (Hamilton & Gifford, 1976). This finding has been replicated under different conditions and is very robust (see for an overview, Hamilton & Sherman, 1989). An important reason for the popularity of this concept lies in its practical implications. The study of the illusion can give us an insight into the processes underlying the formation of social stereotypes and negative attitudes towards minorities in society.

An experiment that shows many of the typical findings in illusory correlation research conducted by McConnell, Sherman and Hamilton (1994, experiment 2) will be used here to illustrate our simulation of the bias. Before proceeding to the simulation, we first discuss the most important empirical measures of the illusion and previous rival explanations.

Evaluative Judgments.   The majority of illusory correlation studies used the same set of measures that were originally introduced by Hamilton and Gifford (1976) and that were also used by McConnell et al. (1994). As noted earlier, in many studies these measures showed an evaluative bias in favor of the majority group.

- *Likability ratings*: McConnell et al. (1994, p. 416) asked their participant to rate "how much they thought they would like members of Group A and Group B" on a 10-point scale ranging from "strong disliking" to "strong liking", and found that group A was liked most.
- *Frequency estimates*: For each group, participants were asked "to estimate how many of [the behaviors] were undesirable" (p. 416). The number of undesirable behaviors performed by minority group B members was overestimated relative to the number of undesirable behaviors performed by majority group A members.
- *Group Assignment*: Participants were given each behavior without group assignment and then had to indicate "whether a member of Group A or Group B performed the action" (p. 416). It was found that disproportionately more undesirable behaviors were attributed to the minority group B than to the majority group A.

Process Measures. The previous measures record the extent of the illusion, but reveal little about the underlying encoding and memory processes that may be responsible for it. In order to explore these processes in more depth, researchers introduced additional process measures. Although the results obtained with these measures are less robust than those obtained

with the traditional evaluative measures, they avoid guessing strategies that may cloud memory measures. The following results have been reported:

- *Free recall*: Participants were instructed "to write down as many of the behaviors as they could recall" (McConnell et al., 1994, p. 416) without receiving any cue about behavior or group. It was found that they remembered disproportionately more undesirable behaviors of minority group B than any other condition. This may imply better encoding and memory of these undesirable minority B behaviors.
- *Assignment latencies*: In this process measure, the latencies in the group assignment task (see above) are recorded. McConnell et al. (1994) found that participants are fastest in assigning undesirable behaviors to the minority group B (but see Klauer & Meiser, 2000). As this effect in group latencies shows the same pattern as the free recall data, it was again interpreted as a result of better encoding and memory of these behaviors.

Prior Theoretical Accounts. What are the theoretical explanations provided for this pervasive bias? The first account of illusory correlation proposed by Hamilton and Gifford (1976) was inspired by Chapman's (1967) original explanation that centered on the distinctiveness or salience of stimuli that form a minority. Hamilton and Gifford (1976) argued that the co-occurrence of two infrequent events, that is, undesirable behaviors from a minority group, are particularly attention getting and distinct, and therefore received more extensive encoding, which in turn leads to greater accessibility in memory. Because in typical illusory correlation experiments undesirable behaviors are a minority, they become especially salient and memorable in the minority group B. This memory advantage of undesirable group B behaviors was assumed the key factor causing the negative group impressions of the minority group B. The distinctiveness-based explanation has gained quite a lot of empirical support (for extensive reviews see Hamilton & Sherman, 1989; Mullen & Johnson, 1990) that was corroborated by recent studies in which higher recall for distinct undesirable minority group was documented (Hamilton, et al., 1985; McConnell et al., 1994; Stroessner et al., 1992). Despite the popularity and empirical support of the distinctiveness account, however, alternative approaches to illusory correlation have been put forward.

According to exemplar models of Smith (1991) and Fiedler (1996), aggregation of more information reduces unsystematic error and so leads to perceptions that are more accurate. As a consequence, for majority group A that contains a large number of behaviors, the difference between desirable and undesirable behaviors is more accurately perceived than for minority group B, where there are less behaviors. Based on these differences, exemplar models predict an illusory correlation bias, that is, more favorable liking of the majority group. Unlike the distinctiveness account, these models posit that unequal frequencies are responsible for the effect, not selective memory. Similarly, the tensor-product model (Kashima et al., 2000) proposes that the encoding and aggregation of unequal frequencies by means of the Hebbian learning algorithm drives the illusion. Thus, Kashima et al. (2000) emphasize encoding rather than retrieval as the basis of the illusion. However, the increased recall for infrequent and undesirable behaviors noted earlier (Hamilton, et al., 1985; Klauer & Meiser, 2000; McConnel et al., 1994; Stroessner et al., 1992) is currently problematic for both the exemplar and tensor-product account, as they do not address this memory advantage.

To resolve the discrepancy between increased evaluation and decreased memory, alternative models have been put forward (e.g., Garcia-Marques & Hamilton, 1996) that emphasize a dual-retrieval process in which likeability and frequency estimates depend on the spontaneous availability or ease of retrieval of the episodic items, while free recall depends on an exhaustive search guided by the number and direction of the links between episodic nodes.

However, such models are strongly limited by the fact that they do not account for the development of group impressions.

     Recurrent Account. Like the tensor-product model, our connectionist account also assumes that illusory correlation is created by differences in sample sizes that affect encoding rather than by memory retrieval differences between behaviors. Because of the acquisition property of the delta algorithm, the prevalence of desirable (relative to undesirable) behavior is more clearly encoded in the evaluative connections for the majority group, so that perceivers have a more positive impression of the majority group compared to the minority group. In addition, increased memory for undesirable minority behaviors is driven by the competition property of the delta algorithm as described earlier.

     Table 2 represents a simplified simulated learning history of a typical illusory correlation experiment as conducted by McConnell et al. (1994, experiment 2).   Each line in the top panel of Table 2 represents a pattern of external activation at a trial that corresponds to a statement presented to a participant. The first two cells represent the group label present in each statement, the next two cells denote the valence of the statement, and the last cells represent episodic nodes reflecting the behavioral information presented [2].

     In the simulation, to measure the traditional evaluative judgments on the groups (i.e., likability ratings, frequency estimations and group assignments), the group nodes were turned on and the resulting activation of the evaluative nodes was read off (denoted by ?, see bottom panel of Table 2). As noted earlier, no additional external activation was provided to the evaluative nodes (or any other "measurement" node) because null activation is a neutral resting activation state that allows an unbiased assessment of the evaluative activation generated directly or indirectly by the group nodes. In particular, we tested the resulting differential activation from the desirable and undesirable node. Although it is also possible that the evaluative nodes are first primed and that this activation then travels to the group nodes, this has little effect on the network's predictions. The reason is that the sample size effect that drives the illusion is largely symmetric over opposing directions of the evaluative connections.

     As discussed earlier, episodic memory can be measured by a group assignment task, preferably by measuring latencies that avoid contamination by guessing strategies or response biases that are driven by evaluative memory. In a group assignment task, behaviors are presented and participants have to indicate as fast as possible by which group member they were performed. To reflect this measure, each episodic node from different sets of behaviors (A+, A-, B+, B-) was activated one at a time (see bottom panel of Table 2). This episodic activation spreads to the group nodes and so determines response times. This testing procedure is based on the assumption that awareness of group membership depends on the crossing of a minimal activation threshold (Cleeremans & Jiménez, 2002). By assuming that the time to spread the activation through the network is proportional to the strength of the connection weights, stronger episodic→group connections will lead to higher group activations and faster crossing of the awareness threshold for group membership [3].

     Results. The 18 "statements" listed in Table 2 were processed by the network for 50 "participants" with different random orders. Figure 5 depicts the mean test activation for all simulated dependent measures, together with the observed likeability and reaction time data from McConnell et al. (1994, exp. 2).   The top panel of the figure shows the results of the simulation of the evaluative measures, together with the likeability ratings from McConnell et al. (1994, exp. 2). The simulation shows that the majority group A received higher evaluative activations than the minority group B, $F(1, 49) = 39.05$, $p<.001$, mirroring the same pattern of the observed data (the perfect fit is exceptional and simply due to the rescaling of the test

activation of the network to the observed data that consists here only of two data points).

The bottom panel depicts the results of the memory simulation together with the observed assignment latencies of McConnell et al. (1994, exp. 2). Although the observed differences between B+ and B- are somewhat underestimated, as predicted, the competition property resulted in stronger episodic connections for minority behaviors, $F(1, 49) = 425.44$, $p<.001$, and undesirable behaviors, $F(1, 49) = 264.28$, $p<.001$. These two main effects indicate that it is not the combination of negative and minority behaviors (i.e., B-) that might drive the illusion as the distinctiveness account would predict, but rather two independent effects of increased memory stemming from two minority categories (behaviors from group B and negative behaviors). As noted earlier, these results also distinguish the present network from alternative exemplar-based and tensor-product models that cannot account for the increased memory for minority groups and undesirable behaviors.

Simulation 2: Expectancy-based Illusory Correlation

Although the differential sample size paradigm of Hamilton and Gifford (1976) represents a very dramatic demonstration of illusory correlation despite the lack of an actual relationship, very often group stereotypes are created as a consequence of existing relationships between attributes and a group. Once such group conceptions are formed, however, these beliefs will bias judgments based on newly acquired information, even if that new information does not contain an actual relationship. Thus, already established stereotypes may produce illusory correlations through the expectations that are associated with a group. Therefore, this type of illusory correlation is termed expectancy-based, in contrast to the Hamilton and Gifford (1976) paradigm that we refer to as size-based.

In an illustration of this expectancy-based illusory correlation, Hamilton and Rose (1980, exp. 1) presented their participants with a series of statements, each of which described a person as a member of an occupational group such as accountants and doctors. In addition, each member was described by two trait-implying adjectives, some of which were stereotypically associated with the group while others were not. For instance, the traits *perfectionist* and *serious* were stereotypical of accountants, and the traits *wealthy* and *attractive* were stereotypical of doctors. All these trait adjectives were presented in descriptions of all occupational groups, so that there was no relationship between occupational group and any particular attribute. Moreover, there were always two members associated with each set of two adjectives, so that sample size was kept constant. Nevertheless, when asked to indicate "how many times each of these adjectives described each occupational group" (p. 835), participants overestimated the frequency of traits that were stereotypical of a group. For instance, they estimated the frequency of perfectionist and serious accountants to be on average 2.7 (while the actual number was 2). In contrast, the frequency of doctors having these traits was estimated to be 2 (which was the actual number).

This finding cannot be explained by differences in sample size in the information set. Apparently, preexisting expectancies about these occupational groups had biased the frequency estimates of co-occurrence. Subsequent studies have replicated these findings (Kim & Baron, 1988; Slusher & Anderson, 1987; Spears, Eiser & van der Pligt, 1987).

Several explanations have been put forward to account for expectancy-based illusory correlation, including facilitated encoding of stereotypical traits or biases at retrieval. We propose a connectionist explanation that builds on the suggestion by Hamilton and Rose (1980) that "an associative basis for an illusory correlation would exist whenever one's previous experiences had resulted in a perceived relationship between two stimulus variables. The perceiver would then have an expectation that the two variables are related" (p. 833).

Specifically, we assume that the bias results from previous experiences with co-occurences of stereotypical traits with an occupational group, and so creates pre-existing stereotypical beliefs that are encoded in stronger weights connecting the stereotypical traits with the group. Consequently, when novel information is presented, the new weight changes resulting from this information are "added" on these prior weights, leading to a stereotypical weight advantage. These stronger weights for stereotypical traits produce the illusory correlation. More generally, the integration of old and new information in a connectionist model by adding weight changes, explains how expectancy-driven biases are created.

We used the same model architecture as depicted in Figure 2 for size-based illusory correlation, with the exception that trait nodes replace the desirability nodes. However, the present simulation is driven by another property of the delta algorithm, the modification of weights derived from old and new information. Specifically, during a pre-experimental phase, the model builds up an association or expectancy about typical traits of each occupational group by presenting 5 trials in which stereotypical traits co-occur with their occupational group (without any episodic information on specific trait adjectives, as this information is most probably lost by the time the experiments starts). Next, during the experimental phase, information is presented that was either consistent or inconsistent with the stereotype, leading to a zero correlation overall. At the end of learning, to simulate frequency estimates which reflect "how many times each of these adjectives described each occupational group" (p. 835), each group is primed and the resulting activation of each trait node is read off (the reverse direction of testing from trait to group nodes works equally well). Because we simulated single traits (without the presence of opposing trait), simulation of the frequency measure was tested by the resulting activation of a single trait only (instead of the usual differential activation).

Results. Like the previous simulation, the network processed all "trait adjectives" for 50 "participants" with different random orders. The mean test activations for the simulated frequency estimates are depicted in Figure 6, together with the observed means for two occupational groups from the first experiment of Hamilton and Rose (1980). As can be seen, the simulation replicates the basic finding that stereotypical traits are overestimated in frequency in comparison with non-stereotypical traits. A within-subjects ANOVA revealed that, like in the original study of Hamilton and Rose (1980), the interaction between Group (accountants versus doctors) and Typicality (typical of accountant versus doctor) reached significance, $F(1, 49) = 5554.86, p<.001$.

Group Differentiation

Several biases and stereotypes in group judgments such as illusory correlation emerge from categorizing people or objects in different groups. A factor that exacerbates the creation of stereotypes is accentuation, or the tendency to exaggerate differences on a feature that determines group categories (Tajfel, 1969). For instance, differences between skin colors are exaggerated between blacks and whites, but are seen as more similar among people belonging to the same racial group. In a classic study, Tajfel and Wilkes (1963) reported that when short and long lines were systematically associated with different categories, the perceived difference between the short and the long lines became more pronounced while similarities of the items within each category were increased (but see Corneille, Klein, Lambert & Judd, 2001). Such accentuation leads to less individuation and hence more stereotypical beliefs about social categories.

Early theories remained vague about the psychological process underlying the accentuation effect. For instance, Tajfel and Wilkes (1963) suggested that the main drive behind

the effect is a desire to maximize predictability. Cognitive explanations have also been offered: Exemplar theories (Fiedler, 1996; Krueger & Clement, 1994) assume that in a correlated condition, attention to the group label of an exemplar leads to the recruitment from memory of more exemplars from the same group, which are then aggregated into a composite evaluation that gives more weight to exemplars of the same group than from other groups. This increases the perceived similarity within groups and difference between groups. The tensor-product model of Kashima et al. (2000) proposes a similar account. Because of the correlation between exemplars and the category, all exemplars of the same category share a common group label, and so become more similar to each other and more different from other groups. In sum, both the exemplar and tensor-product model offer an account of the accentuation effect in terms of the sample size of the group category.

Like the exemplar and tensor-product theories, our recurrent network also offers a sample size account. The idea is that accentuation is produced by the group→attribute connections. Because a correlated condition implies a greater sample size of the co-occurrence of a group and attribute nodes, based on the acquisition property, stronger group→attribute connections will develop. For example, if eight pro-gay articles are all correlated with one newspaper, strong associations will develop between the newspaper source and this attitude position. In contrast, when four pro-gay articles are correlated with one newspaper and another four to another newspaper, the connections of the each of the newspaper sources with the attitude position will be much weaker.

For the group→attribute connections to have any effect on judgment, we assume that when perceivers judge an exemplar, not only the episodic trace but also the newspaper source is activated to some degree. As noted earlier, this assumption was also made by previous exemplar and tensor-product theories. Moreover, recent findings corroborate the idea that accentuation is more likely to emerge when the task is sufficiently complex, suggesting that especially under such conditions participants additionally rely on categorical (i.e., source) information (Corneille et al., 2001; Lambert, Klein & Azzi, 2002). Because of the stronger group→attribute connections in the correlated condition, this leads to accentuation of differences with the other group that would not occur if group labels were not correlated. For example, because the connection between a newspaper source and the pro-gay attitude in the correlated condition is stronger, activating this newspaper node will result in higher activation of the pro-gay attitude node (and almost no effect on the anti-gay node as this newspaper was obviously not correlated with anti-gay articles), leading to increased pro-gay ratings or accentuation. In contrast, because this connection is weaker in the uncorrelated condition, activating the newspaper node will result in relatively weaker activation on the pro-gay attitude node, leading to less pro-gay ratings and loss of accentuation. The reasoning is similar for anti-gay articles and the anti-gay attitude node.

Novel Prediction and Initial Empirical Support

The present account makes a novel prediction that earlier exemplar models (Fiedler, 1996) or the tensor-product model (Kashima et al., 2000) do not make. Given that the effect of acquisition is largely symmetric over the evaluative connections, not only the group→attribute connections should be weaker in the uncorrelated condition than in the correlated condition as described above, but also the attribute→group connections. By the competition property, this should lead to less discounting of the episodic→group connections (just as it was the case for minority groups in illusory correlation). Hence, our recurrent model predicts that the episodic→group connections should be stronger in the uncorrelated than in the correlated condition, leading to better recognition (assignment of source labels).

To verify this prediction, Vanhoomissen, De Haen and Van Overwalle (2001) explored the effect of classification on accentuation of attitudes. Participants were presented with statements reflecting favorable versus unfavorable attitudes towards homosexuality, that came ostensibly from two newspapers (cf. Eiser, 1971). In a correlated condition, the favorable statements were consistently attributed to one paper and the unfavorable statements to another paper. In contrast, in an uncorrelated condition, statements were attributed equally often to each newspaper. After reading the statements, the participants were requested to rate all the statements on an 11-point scale ranging from *very negative* to *very positive*. In line with the accentuation prediction, the difference between favorable and unfavorable statements was accentuated in the correlated condition as compared to the uncorrelated condition.

In addition, a newspaper assignment task was included. Participants read the original statements as well as novel distracter statements (foils) that contained the same material, but differed in their evaluative meaning (i.e., switched from favorable to unfavorable and vice versa), and they had to indicate from which newspaper the statements came or whether it was not presented earlier. The rationale behind the foils was that this would allow unconfounding episodic memory from guessing on the basis of evaluative memory. If the participants were (mis)led by the evaluative meaning of the statements, we would find worsened recognition performance on the foils, in that they would not be sufficiently rejected. However, if the participants were led by their episodic memory of the statements, they should show improved recognition performance on the foils, that is, they should reject them more often.

Our novel recurrent prediction for the recognition task was better (episodic) memory of the foils in the uncorrelated condition than in the correlated condition. Consistent with this prediction, in the recognition task, participants more often rejected distracter foils in the uncorrelated condition than in the correlated condition. This suggests that, compared to the correlated condition, these participants were less often misguided by the evaluative implication of the foils and used their episodic memory for making correct recognition judgments. Conversely, as one would expect, participants in the correlated condition more often accepted the original items, indicating again that they were (in this condition) correctly guided by the evaluative implication of these statements.

Simulation 3: Accentuation

A recurrent implementation of Vanhoomissen et al.'s (2001) accentuation and assignment findings is given in Table 3. We used the same semi-localist encoding of attribute (attitudes) and episodic (articles) information as before. Again, we simulated 50 "participants" with different random orders. To measure accentuation, participants were requested to give an estimate of the attitude position of each stimulus (e.g., how much pro- or anti-gay each statement was). Hence, in the network, we tested for accentuation by cuing each episodic node representing an article as well as its associated newspaper node. (To unconfound source from favorability across the two correlation conditions, we activated only four favorable and four unfavorable articles that consistently came from the same newspaper in the two correlation conditions). The degree to which this activation spreads to the attitude nodes determines the perceived attitude strength of the articles (see bottom panel in Table 3). The best fit with the observed data from Vanhoomissen et al. (2001) was obtained when the newspaper nodes were activated only for .15 rather than the default value (suggesting that belongingness to the newspaper was recruited from memory not to its full degree; the same .15 activation value provided the best fit in simulations of a similar study by Eiser, 1971).

In addition, we measured rejection of the foils in the newspaper assignment task. We assumed that this rejection would follow as a function of the conflicting group activations

associated with (a) the behavior described in the foils and (b) the reversed attitude positions. Therefore, we first tested recognition of the foils, by measuring how participants "falsely" recognized the foils as belonging to the original group. Specifically, we activated the foils by priming each episodic statement together with the reversed attitude position, and read off the resulting activation from the newspaper group nodes (see last two rows in the bottom panel of Table 3). Next, we measured the conflict with the group activation arising from the reversed attitude position. To accomplish this, we did the same test as before except that we only primed the reversed attitude positions and then subtracted the resulting group activation from that obtained for the foils. This difference score reflects the experienced conflict arising from episodic and reversed attitude information. The greater the conflict, the more likely the foil will be recognized and rejected.

Results. Figure 7 shows the simulation results of 50 randomized "participants". As can be seen on the top panel of the figure, the simulation demonstrates a clear accentuation effect in that the perceived attitude positions were more extreme in the correlated condition compared to the uncorrelated condition, and the expected interaction was significant, $F(1, 98) = 121.46$, p<.001. In addition, the bottom panel shows that our novel memory prediction was also supported as episodic memory was higher in the uncorrelated condition than in the correlated condition, $F(1, 98) = 438.83$, p<.001.

This demonstrates that our recurrent network can model accentuation and the associated effect of enhanced memory for uncorrelated attributes. We argue that the network's ability to reproduce the accentuation effect is due to sample size sensitivity of the acquisition property, while enhanced recognition (i.e., assignment) is due to the competition property. Other theories such as the exemplar-based model of Fiedler (1996) and the tensor-product model by Kashima et al. (2000) make the same accentuation prediction, but are silent with respect to enhanced recognition.

Stereotype Change

So far, we have seen how cognitive processes in humans — as modeled by a recurrent network — may shape distorted impressions about groups. The important question then is how might we be able to get rid of these biased impressions? Three tactics for providing stereotype-inconsistent information have been proposed in the literature to counter biased group perceptions (Weber & Crocker, 1983; for an overview, see Hewstone, 1994):

According to the conversion model, extreme group members have an especially strong impact on perceptions of a group as a whole, so that disconfirming behavior of these members is especially likely to change group stereotypes. However, this model has received little empirical support. More evidence was found for the bookkeeping model, which predicts a gradual modification of stereotypes by the additive influence of each piece of disconfirming information. Thus, for instance, more frequent disconfirming information will elicit more changes (Weber & Crocker, 1983). This prediction is in line with the recurrent model, as the acquisition property also predicts that more evidence leads to more extreme judgments (see also e.g. the sample size effects on illusory correlation and accentuation, discussed earlier).

Perhaps the subtyping model has inspired the most promising tactic. This model predicts that extreme group members will be subtyped into subcategories and separated from the rest of the group. This insulates the group from dissenting members, so that the content of the existing group stereotype is preserved. Hence, contrary to the conversion model, this model predicts that the best tactic to change group stereotypes is to distribute disconfirming information among as many group members as possible, so as to avoid subtyping of extreme disconfirmers. Empirical

evidence has generally supported this prediction (Hewstone, Macrae, Griffiths & Milne, 1994; Johnston & Hewstone, 1992; Weber & Crocker, 1983).

For instance, Johnston and Hewstone (1992, exp. 1) provided stereotype-inconsistent information on occupational groups that was either dispersed across many members or concentrated within a few members. When participants were asked how characteristic several stereotype-consistent and inconsistent traits were of the group in general, they showed a strong increase of stereotype-inconsistent traits in the dispersed condition. Frequency estimates of each type of information showed the same pattern, that is, higher estimates of inconsistent information in the dispersed condition. When asked to rate the typicality of the confirmers and disconfirmers in each group, it was found that the disconfirmers were seen as much less typical in the concentrated condition. This suggests that, as predicted by the subtyping model, disconfirmers were probably subcategorized more in the concentrated condition than in the dispersed condition.

Can the recurrent model reproduce these changes? It can, by simulating subtyping through the property of competition. We again assume a semi-localist representation in which not only the trait description is encoded in a stereotype-consistent or inconsistent node, but also the person to whom the trait is attributed. When stereotype-inconsistent information is concentrated in a few members, this implies that after repeated presentation, the exemplar nodes representing these disconfirming members develop their own strong connection with the inconsistent node. (This is less the case for confirming members, because their exemplar→consistent connections are blocked by the strong group→consistent connection.) These strong exemplar→inconsistent connections compete with the group→inconsistent connections, resulting in a discounting of this latter connection. Psychologically, this leads to a decreased impact of inconsistent information on the group as a whole. In addition, because the disconfirming exemplar nodes develop stronger connections with the inconsistent node as noted above, this results in a greater impact of the few disconfirming members on inconsistency ratings, resulting in these members being recognized as more inconsistent compared to the majority (i.e., subtyping).

In contrast, when the stereotype-inconsistent information is dispersed across members, these exemplar nodes do not develop strong connections with the inconsistent node, so that no competition arises with the connections linking the group with the inconsistent traits. Hence, no discounting of the inconsistent information occurs and no subtyping appears. In sum, the connection linking the group with the inconsistent node is more discounted by disconfirming members in the concentrated condition than in the dispersed condition, leading to a conservation of stereotypical perceptions of the group as a whole. In addition, the stronger connections of disconfirming members with the inconsistent node in the concentrated condition results in more subtyping of disconfirming members away from the rest of the group.

Simulation 4: Dispersed versus Concentrated Stereotype-inconsistent Information

Table 4 lists a recurrent implementation of Johnston and Hewstone's study (1992, exp. 1). As can be seen, the network architecture consists of a group node, two trait nodes reflecting stereotype-consistent and inconsistent traits, and several exemplar nodes reflecting individual members. The representation of stereotype-consistent and inconsistent traits as two unitary nodes is similar to the representation in the illusory correlation network (Simulation 1) of behaviors in desirable and undesirable nodes. In contrast to the earlier simulations, however, the exemplar nodes only represent members, and not their behaviors (which were not simulated). To provide the network with prior expectancies on stereotypical beliefs of the group, we provided

10 trials of stereotypical traits in a pre-experimental phase. Next, in the concentrated condition, all disconfirming information was concentrated in the last two group members, whereas in the dispersed condition, disconfirming information appeared in all members except the first two. The overall amount of inconsistent information was identical (i.e., 12) in the two conditions.

Again, we simulated "50 participants" with different random orders. To measure stereotypical beliefs, participants are typically requested to rate to what extent some stereotype-consistent and inconsistent traits describe the group (Johnston & Hewstone, 1992; Johnston, Hewstone, Pendry & Frankish, 1994; Weber & Crocker, 1983). In the network, this was tested by cuing the group node and reading off the resulting activation on the consistent or inconsistent node. We assume that frequency estimates are based on a similar testing procedure (see also illusory correlation, discussed earlier). To measure subtyping, Park, Wolsko and Judd (2001) demonstrated that one of the more valid measures was to request the perceived typicality of confirming and disconfirming group members. In the network, this was tested by activating the two members that were either confirmers or disconfirmers in both conditions, and reading off the resulting trait activation (bottom panel in Table 4).

Results. Figure 8 shows the simulation results of 50 randomized "participants" on the trait ratings (top panel) and the typicality ratings (bottom panel). As can be seen in the top panel, the simulation demonstrates no considerable difference for consistent traits and, more importantly, a substantial effect of discounting of inconsistent traits in the concentrated condition as opposed to the dispersed condition. That is, the inconsistencies were less strongly associated with the group in the concentrated condition than in the dispersed condition, as in Johnston and Hewstone's (1992) study. A between-subjects ANOVA confirmed that the difference between the two conditions was significant for inconsistent traits, $F (1, 98) = 26.29$, p<.001, but not for the consistent traits, $F (1,98) < 1$, *ns*. In addition, the bottom panel shows lower typicality ratings for disconfirmers in the concentrated than in the dispersed condition (and, as one would expect, almost no differences for confirmers). Again, this difference was significant, $F (1, 98) = 1572.17$, p<.001. This suggests that disconfirmers in the concentrated condition are more easily subtyped away from the overall group stereotype.

This simulation demonstrates that a recurrent network can model subtyping. The network's ability to reproduce this effect is due to the property of competition, which allows discounting of inconsistent information concentrated in a few disconfirmers. To be precise, disconfirmers are not discounted, but rather their implications for the whole group are. Other theories such as the exemplar-based model by Fiedler (1996) and the tensor-product model by Kashima et al. (2000) do not posses this property, and hence cannot make this prediction except by adding auxiliary assumptions. For instance, Kashima et al. (2000) assumed that the amount of stereotype change is mediated by the extent to which inconsistent group members are individuated away from the group's resting state (p. 931), a process that was added to the model to incorporate individuation in social judgments. In the recurrent model, such additional individuation process was not necessary, because the results came out naturally from the competition property of the delta algorithm.

Very recently, Queller and Smith (2002) proposed another recurrent connectionist model to model subtyping processes. Although many specifications and parameters of their model differ from our network (i.e., distributed representation, symmetric weights, contrastive Hebbian learning algorithm), the basic architecture and processing mechanisms are very similar. However, a more important difference is that Queller and Smith (2002) focused on the distribution of counterstereotypic information among behaviors rather than persons. That is, their simulations do not reflect whether discrepancies are concentrated among a few members or

dispersed among most, but instead reflect a difference between moderate and extreme disconfirming members, differing only in the number of counterstereotypic behaviors. This variation adequately reflects their own experiment with human subjects (experiment 3; see also Weber & Crocker, 1983, experiment 2), and certainly has merit because it points to other mechanisms underlying subtyping.

Based on their simulations, Queller and Smith (2002) concluded that earlier explanations of subtyping are not important. However, it is invalid to generalize these conclusions to the more typical case of subtyping when inconsistencies are concentrated within a few members. Instead, as we claimed earlier, our simulations confirm that in order to change stereotypes of a group, it is essential that discrepant members are still seen as a member of the group, so that the link between group membership and counterstereotypic attributes is not weakened. It is interesting to note that in spite of these differences, our network can reproduce Queller and Smith's (2002) simulation showing that subtyping is reduced when counterstereotypic information is presented throughout with stereotypic information (e.g., when learning about a novel unknown group), instead of after an initial stereotypic phase (e.g., when unlearning stereotypes of a known group for which one has already developed strong stereotypes).

Moderating Factors

The present network can simulate other findings in the literature that examined the effects of several moderating variables on subtyping:

- *Sample Size*. Weber and Crocker (1983, exp. 1 & 2) reported more stereotype change in the dispersed condition when more inconsistent information was provided. The model explains this finding by sample size differences. A growing sample size leads to more inconsistency information being incorporated in the group schema for the dispersed condition, but being discounted and subtyped in the concentrated condition. This can be simulated in the network by increasing the number of inconsistent trials (e.g., by doubling their frequency).

- *Individual Members*. Gurwitz and Dodge (1977) reported that, in contrast to group judgments, estimates of individual members were seen as less stereotypical in the concentrated than in the dispersed condition. However, Weber and Crocker (1983) did not replicate this finding as they found the same pattern of results for individual members as for the whole group. In line with their finding, our simulation also predicts the same overall pattern for individual group members as for group judgments (by placing 1s on the member nodes instead of on the group node, see first two lines in the bottom panel of Table 4).

- *Expectancy*. Johnston et al. (1994, exp. 3) documented more stereotypical ratings when stereotypical beliefs about groups were made explicit (high expectancy) than when they were not made explicit (low expectancy). In addition, in what may appear a ceiling effect, she also found less change when expectancy was high rather than low. To reproduce Johnston et al.'s (1994) findings, low expectancy can be simulated in the recurrent network by reducing the pre-experimental trials (e.g., 2) in comparison with the high expectancy condition (e.g., 10).

Perceived Group Variability

Thus far, we discussed how categorization between groups may distort how we perceive the central tendency of a group attribute (e.g., likeability, attitude, stereotype). However, the perceived homogeneity or variability of people is also strongly affected by group categorization.

On the basis of available evidence, Dijksterhuis and van Knippenberg (1999) concluded that "variability judgments are quite accurate (in the sense that they reflect the actual stimulus variation quite well) and are being updated continuously" (p. 529). This is consistent with a connectionist approach in which group characteristics such as variability are updated on line. The concept of group variability is important, because high variability implies inconsistencies in the relationship between a group and some attribute, and this may help to dilute or change undesired group stereotypes. However, in contrast to Dijksterhuis and van Knippenberg's (1999) claim, research has also documented a number of shortcomings and biases in perceived group variability. Before we turn to these biases, we first briefly discuss how variability is measured in prior research and modeled in our network.

Simulation of Group Variability

A crucial question is how group variability is measured. Research addressing this issue has used a plethora of measures. Park and Judd (1990) analyzed these different measures and found that two independent constructs account for perceived variability. The first construct can be conceived as the dispersion of group members around the mean of one attribute, while the second construct reflects the degree to which the group as a whole is seen stereotypically. We focus here on the first construct, involving measures of perceived dispersion. Park and Judd (1990) reported that the *perceived range* measure was the most valid of group variability. Other measures inspired by an exemplar approach (Linville, Fischer & Salovey, 1989) known as "perceived variability", "probability of differentiation", or direct ratings of perceived similarity seemed less valid.

In a recurrent network, variability can be simulated by an approximation of the range measure. In this measure, participants are given a bipolar rating scale spanning the low to high ends of the attribute and asked to indicate where the most extreme (opposite) members would fall (Simon & Brown, 1987). To answer this question, we suggest that participants consider the group and estimate to what extent this group implicates each opposing attribute. This is simulated during testing by priming the group node and reading off the resulting activation on the attribute nodes, just like in a central tendency measure. However, to measure the distance or range between the attributes, these two resulting activations are then summed, rather than subtracted as in a central tendency measure. (The reverse direction of testing by which first the two opposing attributes are primed and then the activation of the group node is read off gives very similar results) [4].

We chose the implementation of the range measure for several reasons. First, it is the most valid measure of group variability (Park & Judd, 1990) and it reflects actual judgments by participants (of range) rather than experimenter-based calculations (of variance). Second, it is cognitively least demanding because it makes use of information that is already available in memory under the form of group→attribute connections, and is thus more likely recruited spontaneously when judging group variability. Third, it is consistent with the finding (Park & Hastie, 1987) that estimates of variance are constructed and stored on-line rather than from retrieved exemplars, as the group→attribute connections on which our range measure is based are developed during learning (using the acquisition property).

To illustrate our implementation of variability as range measure, we simulated an exemplary case. In this simulation, we wanted to demonstrate that variability is sensitive to sample size. Therefore, variability was created by taking for each block of four trials, three group members that possessed the high end of an attribute and only one member that possessed the low end. According to the incremental acquisition property of the delta learning algorithm,

given its greater sample size, asymptote should be reached more quickly for the high end of the attribute than for the low end. To demonstrate this, we measured the results separately for the high and low extreme of the attribute. A direct measure of variability can be obtained by summing the two extremes of the attribute.

The results are depicted in Figure 9. As expected, the central tendency of the high extreme of the attribute approached asymptote much more quickly than the low extreme. This is due to sample size differences, as there are three times more members placed on the high extreme than on the low extreme. However, as more information is provided, the high and low extreme are more spread apart, resulting in more variability. This illustration suggests that the variability measure is susceptible to sample size. That is, when little information is available on group members, the variability of the group is perceived as low. The more information is available, the more the group is seen as heterogeneous until a maximum variability is attained that depends on the spread between the central tendencies of both extremes.

Simulation 5: Outgroup and Ingroup Homogeneity

In group perception, there is a pervasive tendency to perceive an outgroup as less variable than an ingroup, a bias known as the outgroup homogeneity effect (Linville, Fisher & Salovey, 1989; Messick & Mackie, 1989). Research revealed that outgroup homogeneity is related to the fact that perceivers are more familiar with the ingroup and therefore form a more differentiated impression on the ingroup compared to an outgroup (Linville, Fisher & Salovey, 1989). This explanation is also supported by the finding that ingroup heterogeneity is larger for real and enduring groups where everyone knows each other very well, than for artificial and laboratory-created groups (Mullen and Hu, 1989). In line with this explanation, many researchers provided an exemplar-based account of this effect (Fiedler, 1996; Fiedler, Kemmelmeier & Freytag, 1999; Hamilton & Trollier, 1986; Linville, Fisher & Salovey, 1989; Linville & Fisher, 1993; Park & Judd, 1990). Because perceivers have a richer knowledge base of the ingroup, they tend to recruit more exemplar information from memory about the ingroup than the outgroup, leading to more differentiated ingroup judgments.

Our connectionist approach makes a similar prediction as the exemplar approach. Because of the more extensive contact with one's ingroup, perceivers sample more information on the ingroup, leading to more differentiated views of the ingroup. However, contrary to exemplar theories, the connectionist approach assumes that the effect of sample size occurs at encoding rather than retrieval.

Clear support for the sample size account of outgroup homogeneity comes from the finding that the bias can be reversed when the ingroup is not a majority. Under these conditions, the variability of the ingroup is perceived as much smaller than that of the outgroup (Simon & Brown, 1987; Simon & Pettigrew, 1990; Simon & Hamilton, 1994; see for an overview Mullen & Hu, 1989). In a well-known experiment by Simon and Brown (1987, exp. 1), children were arbitrarily assigned to one of two groups (blue or green) depending on their capacity to correctly categorize blue or green colors. Then they were given information on the number of children in each group, indicating that the ingroup was either a minority, a majority or equal in number to the outgroup. Finally, they were asked to estimate the two scale values that would bracket the values of all individuals in each group (i.e., range measure) in their ability to perceive blue and green colors. The results demonstrated that ingroup variability was highest when the ingroup was not a minority (either a majority or equal), and outgroup variability was highest when the ingroup was a minority. This finding was reproduced in the next simulation.

A simplified simulation of the Simon and Brown's (1987, exp. 1) experiment is listed in Table 5. The network consists of an ingroup and an outgroup node, two nodes reflecting the

high and low extremes of the attribute (e.g., good or bad in perceiving blue) and several episodic nodes. As can be seen, some variability in the group was introduced by varying the degree to which members had one of the attributes, that is, by varying the attribute node activation between 0 and +1, including intermediate values of 0.5 and 0.8.   Importantly, to reflect our assumption that perceivers typically have more information on the ingroup than on the outgroup, the ingroup was described by 8 behaviors and the outgroup by 4 behaviors. In contrast, to simulate ingroup homogeneity due to the ingroup being a minority, we simply reversed the group labels so that the ingroup had 4 behaviors and the outgroup 8 behaviors. Perceived group range was tested by activating the group node and reading off the (summed) activation of the high and low attributes, as explained earlier.

> *Results*. The network was run with 50 "participants" with a different random order. As can be seen in Figure 10, the simulation produced a larger variability for the ingroup compared to the outgroup when more information on the ingroup is available (non-minority), thus successfully replicating the outgroup homogeneity effect. In contrast, when the ingroup was a minority, the effect was reversed just as in Simon and Brown (1987, exp. 1). An ANOVA with Ingroup Size as a between-subjects factor and Group (ingroup versus outgroup) as a within-subjects factor confirmed that the interaction was significant, $F(1, 98) = 2110.31$, $p < .001$.

Fit and Model Comparisons

A summary of the simulations that we have reported together with the major property responsible for generating the group biases can be found in Table 1. All simulations replicated the empirical data reasonably well. This can also be verified in Table 6 where the correlations between simulated and observed data are listed. However, it is possible that this fit is due to some procedural choices of the simulations rather than conceptual validity. To demonstrate that changes in these choices generally do not invalidate our simulations, we explore a number of issues, including the localist versus distributed encoding of concepts, and the specific recurrent network used versus a feedforward network. In addition, we will also briefly discuss major differences with other relevant models.

Distributed Coding

The first issue is whether the nodes in the auto-associative architecture encode localist or distributed features. Localist features reflect "symbolic" pieces of information, that is, each node represents a concrete concept. In contrast, in a distributed encoding, a concept is represented by a *pattern* of activation across an array of nodes, none of which reflect a symbolic concept but rather some sub-symbolic micro-feature of it (Thorpe, 1994). Moreover, distributed coding usually implies an overlap of the concepts' representations (i.e. an overlap of pattern activations coding for different concepts). Although we used a localist encoding scheme to facilitate our introduction to the most important connectionist processing mechanisms underlying group biases, we admit that localist encoding is far from realistic. Unlike distributed coding, it implies that each concept is stored in a single processing unit and, except for explicit differing levels of activation, is always perceived in the same manner without noise. This may limit the model's capacity to simulate properties like pattern completion, generalization, and graceful degradation.

For instance, in the semi-localist encoding of our simulations the implied attributes in the statements were directly coded as given, such as whether the behavior was desirable or undesirable, whether the attitude was favorable or unfavorable, and so on. However, participants were not literally told that the statements had these attributes. Therefore, this material is more realistically represented by a distributed encoding scheme, where attribute information is

embedded in a pattern of noisy activations that the recurrent network must abstract from these patterns, just like real participants must do. Given the advantages of distributed coding, is it possible to replicate our localist simulations with a distributed representation?

To address this question, we ran all simulations with a distributed encoding scheme in which each concept was represented by five nodes that each reflects some micro-feature of the concept. We maintained the same level of overlap between concepts that was already introduced in the semi-localist encoding, that is, the overlap consisted of the attribute nodes shared by the exemplars. Although it would perhaps be more realistic to add even more overlap between concepts, this was not done here because that would require ad-hoc assumptions on how much additional differential pairwise overlap there should be between different concepts. We also added random noise to the activation of theses nodes to simulate the imperfect conditions of perception (see Table 6 for details). All simulations were run with 50 "participants" with different distributed representations and random noise for each participant. As can be seen, all distributed simulations attained a good fit to the data and, in all cases, the relevant pattern of results from the localist simulations was reproduced. These findings suggest that the underlying principles and mechanisms that we put forward as being responsible for the major simulation results can be obtained to the same degree not only in the more contrived context of a localist encoding, but also in a more realistic context of a distributed encoding.

Feedforward Model

We claimed earlier that the connections between, attributes, exemplars and group nodes were presumably most responsible for replicating the phenomena of interest. This claim can be partly tested by using a feedforward network model, in which only the feedforward connections from attributes and exemplars to the group play a role (i.e., the upward connections in Figure 2). However, this leaves out the important lateral connections between attribute and episodic nodes, such as the ones involved in the attitude ratings of accentuation (between attitude positions and exemplary statements; Simulation 3) and the typicality ratings of stereotype change (between traits and specific group members; Simulation 4). Thus, except for these two latter cases, we expect a feedforward network to do about equally well as the auto-associative network. To explore this, we ran all simulations with a feedforward pattern associator (McClelland & Rumelhart, 1988) that consists only of feedforward connections (with additional backward spreading of activation from the group node during testing if necessary; see Van Overwalle, 1998). As can be seen in Table 6, for all simulations except those mentioned above, a feedforward architecture did almost equally well as the original simulations. This confirms that feedforward connections are crucial to reproduce many phenomena in group bias. Nevertheless, it is necessary to incorporate lateral connections of a recurrent network to explain all findings of interest.

Non-linear Recurrent Model

We also claimed earlier that a recurrent model with a *linear* updating activation function and a *single internal updating* cycle (for collecting the internal activation from related nodes) was sufficient for reproducing the group biases. This contrasts with other researchers who used a non-linear activation updating function and more internal cycles (McClelland & Rumelhart, 1986; Smith & DeCoster, 1998; Read & Montoya, 1999). Cycling in a recurrent network has some advantages. For instance, it would allow measuring response latencies in an alternative manner by the number of cycles needed to converge on a stable response. (Recall that we simply assumed that the strength of the connection is proportional to the time to spread the activation). Are such activation specifications necessary? To answer this question, we ran all our

simulations with a non-linear activation function and 9 internal cycles (or 10 cycles in total) [5]. Our model specifications were identical to those of Read and Montoya (1999; see also McClelland & Rumelhart, 1988, p. 168—169).

As can be seen from Table 6, although the non-linear model yielded an adequate fit, most simulations did not improve substantially the fit compared to the original simulations. This suggests that the present linear activation update algorithm with a single internal cycle is sufficient for simulating many phenomena in group judgments. This should not come as a surprise. In recurrent simulations of other issues, such as the formation of semantic concepts, multiple internal cycles were useful to perform "cleanup" in the network so that the weights between, for instance, a perceptual and conceptual level of representation were forced to eventually settle into representations that had pre-established conceptual meaning (McLeod, Plunkett, & Rolls, 1998). Such a distinction between perceptual and conceptual levels was not made here, and, as a result, multiple internal cycles seem unnecessary. Nevertheless, non-linear recurrent activation made it possible to simulate accentuation without providing external activation to the source nodes. Whether doing away with this external activation might better reflect real psychological processes is unclear, because research has shown that accentuation does not always occur, and depends on reliance to source categories when the task is ambiguous (Corneille et al., 2001; Lambert, Klein & Azzi, 2002).

Perhaps more importantly, the non-linear activation algorithm tends to abolish the effects of competition in the memory and latency measures (Simulations 1 & 3). The reason is that the non-linear updating algorithm forces the activations automatically to the +1 and -1 default levels. Hence, if two features are activated together and overpredict a category, then the overly high output activation of the category tends to restore to the normal +1 ceiling level. This reduces discounting of the connections with the features.

Exemplar and Tensor-Product Models

As mentioned in the introduction, there are a number of differences between the present recurrent network and exemplar (Fiedler, 1996; Fiedler, Kemmelmeier & Freytag, 1999; Smith, 1991) and tensor-product (Kashima et al., 2000) models. These differences allow the recurrent model to explain more biases in a more parsimonious manner with less assumptions [6]. We discuss these differences in function of the properties that create the biases in the recurrent model:

- *Acquisition*. Exemplar (Fiedler, 1996; Smith, 1991) and tensor-product (Kashima et al., 2000) models explain many group biases by aggregation over samples of different sizes as does our connectionist approach. Consequently, these models can explain biases such as illusory correlation, accentuation, and group homogeneity (Simulations 1, 3 & 5). However, in exemplar models, sample size differences appear only when noise in perception and encoding is assumed, while this assumption is unnecessary in our and Kashima et al.'s (2000) connectionist approach. The reason is that aggregation in a connectionist network is performed during encoding by a learning algorithm that is in itself sensitive to sample size. However, these differences are minor because noise and information loss seem quite plausible, also from a neuropsychological perspective (and they were also used in our distributed simulations).

- *Competition*. Perhaps the most important limitation of exemplar and tensor-product models is that the competition property is absent. The reason is that aggregation in these models is unbounded and has no asymptote. Multiple inputs do not compete against each other for weight but add to the aggregated output in equal amounts. Although these

models often use some sort of a normalization function that limits the overall activation (Fiedler, Kemmelmeier & Freytag, 1999, p. 12; Kashima et al., 2000, p. 918), as we understand it, this function has a global effect that does not cause competition between the summed activation received from multiple inputs like in the delta error-reducing algorithm. Consequently, these models cannot explain enhanced memory for minority behaviors (in illusory correlation) or for uncorrelated conditions (in accentuation), or the effect of dispersed versus concentrated distribution in stereotype change (Simulations 1, 3 & 4). These biases were not discussed in the exemplar models of Fiedler (1996; Fiedler, Kemmelmeier & Freytag, 1999) and Smith (1991; Smith & Zárate, 1992). We see no immediate remedy for the lack of competition in exemplar models. In the tensor-product model, these biases were explained by individuation processes that require additional elaboration and controlled processing. As stated by Kashima et al. (2000), "the individuation process involved in stereotype change and group differentiation was explained in terms of the construction of a person representation. Some mechanism is needed to control…the construction process for a person representation" (p. 935). In contrast, our model assumes that these effects are a natural consequence of the competition property of the delta learning algorithm without any need for a control mechanism. The tensor-product model can avoid such controlled processes by adopting an error-reducing learning algorithm such as the delta algorithm.

- *Group Variation*. All previous approaches modeled mostly the central tendency of the attribute (e.g., liking) in group perception, and did not address variability in stereotyping, with the exception of Fiedler, Kemmelmeier and Freytag (1999). In this approach, variability was measured by cuing memory with a gradually degraded pattern of activation reflecting the ideal attribute features, to instantiate the different scale points spanning the high to low ends of the attribute (Fiedler, Kemmelmeier & Freytag, 1999, p. 15). In our approach, group variation was based on the range measure and modeled by adding the aggregates of the two opposite attributes of the groups (rather than differentiating between them like in central tendency measures). Both approaches are able to model ingroup-outgroup homogeneity (Simulation 5). However, our approach seems preferable because it appears simpler and more direct by using existing memory traces and because it is based on the more reliable range measure (Park & Judd, 1990).

In summary, the present model seems to be better equipped to deal with a number of important issues in group judgments. However, this does not deny the merits of earlier alternative models. In particular, Fiedler's (1996) exemplar model has great historical and conceptual value, as it was the first to point out that simple aggregation processes could explain most basic effects of group biases. It was also an important inspiration in developing our connectionist network model. In addition, these models, and particularly the tensor-product model (Kashima et al., 2000), could be more adequate on other issues or simulations for which they were originally designed.

General Discussion

The simulations in this article illustrate that a recurrent connectionist model is able to account for biases and shortcomings in judgments about groups under diverse conditions. The perspective presented here offers a novel view on how perceivers process social information, by describing how knowledge structures are learned through the development of associations between social concepts. This clearly distinguishes it from earlier associationist approaches that used static networks (with non-adjustable links) to represent logical relationships or constraints

between concepts (Kunda & Thagard, 1996; Read & Marcus-Newhall, 1993; Read & Miller, 1998; Shultz & Lepper, 1996). An important advantage of the dynamic and adaptive nature of the present recurrent network compared to these previous models as well as other connectionist models like the tensor-product model (Kashima et al., 2000) is its computational power. This power comes from the delta learning algorithm that generates a number of important emergent properties responsible for a wide range of group biases (see Table 1).

The acquisition property accounted for sample size effects in (the evaluative bias of) illusory correlation, accentuation of group differences and group homogeneity. The competition property accounted for decreased accessibility and lower recall of frequent events in illusory correlation and correlated exemplars in accentuation, and the greater discounting of inconsistent information concentrated in a few group members. Our emphasis on these sometimes neglected properties of the delta algorithm distinguished the present approach from other implementations of the auto-associator (McClelland & Rumelhart, 1988; Smith & DeCoster, 1998) that used properties related to distributed representation (e.g. pattern completion, generalization) to explore cognition. It is also different from other distributed connectionist models of group processes (Kashima, Woolcock & Kashima, 2000) that use the Hebbian learning rule that is unable to reproduce the competition property.

We also presented unique predictions by the recurrent model on illusory correlation and accentuation. Some of these predictions received already some initial evidence. Van Rooy (2001) demonstrated in a series of experiments that the typical illusory correlation in likability ratings and other measures of group evaluation are exacerbated given an increasing smaller sample size, and more importantly, that this effect occurs even in the absence of undesirable minority behaviors (see also Shavitt et al., 1999). This poses clear problems for competing models of illusory correlation such as the distinctiveness account that situate the origin of illusory correlation at an enhanced memory of these infrequent undesirable behaviors (Hamilton & Gifford, 1976; McConnell et al., 1994). Other novel findings demonstrated that memory is enhanced for undesirable behaviors (Van Rooy, 2001) as well as for behaviors that are uncorrelated with a group attribute (Vanhoomissen, De Haen & Van Overwalle, 2001). These results are problematic for exemplar-based approaches (Fiedler 1991, 1996; Fiedler et al., 1993, Fiedler, Kemmelmeier & Freytag, 1999; Smith, 1991) that claim that impaired — rather than increased — information aggregation of rare events is the key factor of illusory correlation and other group biases.

By bringing together biases from traditionally different fields of group research, the presented connectionist approach can contribute to a more parsimonious theory of biases in judgments in several ways. First, simply integrating these findings in this manner, invites for looking at possible further parallels between them. Second, a connectionist approach makes predictions at a more precise level of detail than these previous approaches. Third, Van Overwalle and colleagues (Van Overwalle & Labiouse, 2002; Van Overwalle, Siebler & Labiouse, 2002) — using the same network model — demonstrated that this approach was also able to account for many phenomena in social cognition, including categorization, person impression, assimilation, generalization and contrast, causal attribution, and attitude formation (see also Read & Montoya, 1999; Smith & DeCoster, 1998). These authors also reported that the recurrent model with delta algorithm integrates earlier algebraic theories of impression formation (Anderson, 1981), causal attribution (Cheng & Novick, 1992) and attitude formation (Ajzen, 1991). In addition, our recurrent network parallels basic associative learning principles applied in a growing tradition of studies using associationist theories to human learning (for reviews, see Shanks, 1995; Van Overwalle & Van Rooy, 1998). The revival of associative learning models is largely due to the development of models using error-correcting learning

mechanisms such as the delta algorithm, which has been used widely in the connectionist literature (McClelland & Rumelhart, 1986). In sum, the present connectionist approach places group biases in the wider perspective of the larger field of learning and cognition.

Connectionist models paint a different picture of information processing compared to many earlier models in cognition. They describe the ability of humans to dynamically adjust associations between concepts (groups, attitudes, behaviors …) in a variety of settings (e.g. social, personal). In particular, they assume that automatic and local updating algorithms update these associations, requiring little conscious effort or awareness and without the necessary control of a supervisory device such a central executive. Hence, the connectionist approach provides an answer as to how we are able to form quick impressions of social agents effortless in the rush of everyday life (Bargh, 1996). This is in line with research on stereotyping that shows that prejudiced responses often occur on implicit measures that participants have limited conscious control over (Greenwald & Banaji, 1995; Whitney, Davis & Waring, 1994). This distinguishes the present recurrent approach also from the tensor-product model of Kashima et al. (2000) that, as the authors admitted themselves, "cannot do away with a control mechanism [to explain]...the individuation process involved in stereotype change and group differentiation" (p. 935). In our model, it is assumed that some control could occur, for instance, at the time the information is integrated to produce an explicit answer or judgment.

In addition, it is important to stress that the connectionist learning process on groups is not inherently biased. Many earlier theories of cognition suggested that to cope with the strong demands of the environment, human perceivers resort to biased processes including heuristics (Tversky & Kahneman, 1974), selective attention (Hamilton, 1981), over-generalizations (Tajfel & Wilkes, 1963) and so on. In contrast, within the current framework, this learning process is seen as essentially unbiased. For reasons of evolutionary survival, humans should be capable at detecting at least simple relationships between features in their environment (Wasserman, Elek, Chatlosh & Baker, 1993). Biases arise mainly because of lack or abundance of evidence (e.g. sample size effect), competition between different types of (evaluative versus episodic) information, or instructions and motivational factors that direct the perceiver's attention toward or away from some particular information.

Limitations and Directions for Future Research

While we believe we have shown that a connectionist framework can potentially provide a parsimonious account of a number of disparate phenomena in group judgment, we are not suggesting that this is the only valid means of modeling cognitive phenomena. On the contrary, we defend a multiple-view position in which connectionism would play a key role but would co-exist alongside other viewpoints. We think that a strict neurological reductionism is untenable, especially in personality and social psychology, where it is difficult to see how one could develop a connectionist model of high-level abstract concepts such as existing personality differences, motivation, love, and violence, which obviously remain far beyond the current scope of connectionist modeling.

The strong overlap in the basic architecture and learning algorithm of the present recurrent model of group biases with similar models of social cognition in general (Read & Montoya, 1999; Smith & DeCoster, 1998; Van Overwalle & Siebler, 2002; Van Overwalle et al., 2002) opens a lot of interesting avenues for future research. One such topic might be the differences between group and person perception, which is now a topic of increasing interest (e.g., Hamilton & Sherman, 1996; Welbourne, 1999). Within the recurrent framework, group and person perception are based on the same learning process during which perceivers form on-line connections between features (traits, characteristics) and targets (individuals, groups). We

suggest that differences between group and person perception arise because information concerning individuals (or groups perceived as highly entitative) directs attention to general attributes in the information that might reveal the presence of a trait, while less entitative groups invite less to search for such consistencies (Hamilton & Sherman, 1996; Wyer & Srull, 1989). This increased attention can be implemented in the model by higher activation levels. Because of the raised activation of general regularities such as evaluative meaning and social categories, the recurrent model predicts greater speed of learning of general attribute information and, as a consequence, more discounting of episodic traces. This might result, respectively, in weaker illusory correlations for individuals than for groups, but in worse memory for specific episodes related to individuals than to groups.

Our model can be very flexibly applied to accommodate other relevant findings. Differential attention to some aspects of social information at the expense of other aspects may be relevant for other moderating factors of illusory correlation. For instance, decreased activation and a resulting decrease of learning may explain loss of illusory correlation under increased or incongruent mood (Mackie et al., 1989; Kim & Baron, 1988; Stroessner, Hamilton & Mackie, 1992) or when one's attitude position is already consistent with a majority (Spears, van der Pligt & Eiser, 1985). Conversely, increased activation may explain increases in perceived variability of a group, such as when low status members set themselves apart from a group (Doosje, Ellemers & Spears, 1995). However, as noted earlier, the mechanism that produces attentional differences is not modeled in the present network, and presumably requires the inclusion of additional modules in the network dealing with controlled processes. Even without additional activation assumptions, our model can produce other biases such as Simpson's paradox (Fiedler, Walther, Freytag & Stryczek, 2002; Meiser & Hewstone, 2002).

Another direction for future research might be an integration of research on group stereotype change and attitude change. Typically, these two research topics have been conducted almost independently. Group research has typically emphasized immutable characteristics like race and gender or artificial categorizations like groups A & B. In contrast, attitude research often focuses on thematic issues that unite or divide people in real-life groups. There is considerable research inspired by dual-process models of attitude formation (Chaiken, 1980; Petty & Cacioppo, 1986) that describe how the content of arguments as well as other contextual information may help to change people's attitudes. This research, however, has often neglected the robust finding in group stereotype research that inconsistent information received from a few group members is less effective in changing stereotypical beliefs than that same information received from many (Weber & Crocker, 1983). The present recurrent model was capable to model both processes of distributed inconsistent information (Simulation 4) as well as attitude change (Van Overwalle, Siebler & Labiouse, 2002). Perhaps, by taking a similar integrative approach, social research might become more successful in changing people's stereotypes and attitudes with respect to devalued minority groups in society.

References

Adolphs, R., & Damasio, A. (2001). The interaction of affect and cognition: A neurobiological perspective. In J.P. Forgas (Ed.). *Handbook of affect and social cognition* (pp. 27-49). Mahwah, NJ: Lawrence Erlbaum Associates.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179—211.

Anderson, N. H. (1967). Averaging model analysis of set size effect in impression formation. *Journal of Experimental Psychology, 75*, 158—165.

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Baker, A. G., Berbier, M. W., & Vallée-Tourangeau, F. (1989). Judgments of a 2 x 2 contingency table: sequential processing and the learning curve. *The Quarterly Journal of Experimental Psychology, 41B*, 65—97.

Bargh, J.A. (1996). Automaticity in Social Psychology. In E.T. Higgins & A.W. Kruglanski (Eds.), *Social Psychology: Handbook of basic principles.* (pp. 169-183). New York: Guilford.

Barrow, J.D. (1992). *Pi in the sky: Counting, thinking and being.* Penguin Books.

Biernat, M. & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology, 66*, 5—20.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*, 752—766.

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior, 6,* 151-155.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99*, 365—382.

Cleeremans, A. & Jiménez, L. (2002). Implicit learning and consciousness: A graded, dynamic perspective. In R. M. French & A. Cleeremans (Eds.) *Implicit Learning and Consciousness: An empirical, philosophical and computational consensus in the making* (pp. 1—40). East Sussex, UK: Psychology Press.

Corneille, O., Klein, O., Lambert, S., & Judd, C. M. (2001). Don't muddle up inches and centimeters: Evidence for the role of task complexity in obtaining intercategory accentuation effects with unidimensional estimates. *Psychological Sciences*, in press.

Dijksterhuis, A., van Knippenberg, A. (1999). On the parameters of associative strength: Central tendency and variability as determinants of stereotype accessibility. *Personality and Social Psychology Bulletin, 25,* 527-536.

Doosje, B., Ellemers, N., & Spears, R. (1995). Perceived ingroup variability as a function of group status and identification. *Journal of Social Experimental Psychology, 31*, 410-436.

Eagly, A. H. & Chaiken, S. (1993). *The psychology of Attitudes.* San Diego, CA: Harcourt Brace.

Ebbesen, E. B., & Bowers, R. J. (1974) Proportion of risky to conservative arguments in a group discussion and choice shifts. *Journal of Personality and Social Psychology, 29*, 316—327.

Eiser, J.R. (1971). Enhancement of contrast in the absolute judgment of attitude statements. *Journal of Personality and Social Psychology, 17*, 1-10.

Estes, W. K., Campbell, J. A., Hatsopoulos N. & Hurwitz, J. B. (1989). Base-rate effects in

category learning : A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology, Learning, Memory and Cognition, 15*, 556-571.

Fiedler, K. (1991). The tricky nature of skewed frequency tables: An Information loss account of distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology, 60,* 24-36.

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multi-cue environments. *Psychological Review, 103,* 193-214.

Fiedler, K., Kemmelmeier, M. & Freytag, P. (1999). Explaining asymmetric intergroup judgments through differential aggregation: Computer simulations and some new evidence. *European Review of Social Psychology*, Vol. 10 (pp. 1—40).

Fiedler, K., Russer, S., & Gramm, K. (1993). Illusory correlation and memory performance. *Journal of Experimental Social Psychology, 29,* 111-136.

Fiedler, K., Walther, E. & Nickel, S. (1999). The auto-verification of social hypotheses: Stereotyping and the power of sample size. *Journal of Personality and Social Psychology, 77*, 5-18.

Fiedler, K., Walther, E., Freytag, P., & Stryczek, E. (2002). Playing mating games in foreign cultures: A conceptual framework and an experimental paradigm for inductive trivariate inference. Journal of Experimental Social Psychology, 38, 14—30.

Garcia-Marques, L. & Hamilton, D.L. (1996). Resolving the apparent discrepancy between the incongruency effect and the expectancy-based illusory correlation effect: The TRAP model. *Journal of Personality and Social Psychology, 71,* 845-860.

Gluck, M. A. & Bower, G. H. (1988). From conditioning to category learning : An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227-247.

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

Gurwitz, S. B., & Dodge, K. A. (1977). Effects of confirmations and disconfirmations on stereotype-based attributions. *Journal of Personality and Social Psychology, 35*, 495—500.

Hamilton, D. L., Rose, T. L. (1980). Illusory correlation and the maintenance of stereotypic beliefs. *Journal of Personality and Social Psychology, 39*, 832 845.

Hamilton, D.L. (1981). *Cognitive processes in stereotyping and intergroup behavior.* Hillsdale, NJ: Erlbaum.

Hamilton, D.L., & Gifford, R.K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology, 12,* 392-407.

Hamilton, D.L., & Sherman, J.W. (1989). Illusory correlations: Implications for stereotype theory and research. In D. Bar-Tal, C. F. Graumann, A. W. Kruglanski & W. Stroebe (Eds.), *Stereotype and Prejudice: Changing conceptions* (pp. 55-92). New York: Springer.

Hamilton, D.L., & Trollier, T.K. (1986). Stereotypes and stereotyping: An overview of the cognitive approach. In J.F. Dovidio & S.L. Gaertner (Eds.). *Prejudice, discrimination and racism* (pp 63-127). London: Academic Press.

Hamilton, D.L., Dugan, P.M., & Trollier, T.K. (1985). The formation of stereotypic beliefs: Further evidence for distinctiveness-based illusory correlation. *Journal of Personality and Social Psychology, 48*, 5-17.

Hansen, R. D. & Hall, C. A. (1985). Discounting and augmenting facilitative and inhibitory forces: the winner takes all. *Journal of Personality and Social Psychology, 49*, 1482--

1493.

Hastie, R. & Kumar, P. A. (1979) Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology, 37*, 25—38.

Hastie, R. (1980). Memory for behavioral information that confirms or contradicts a personality impression. In R. Hastie, T. M. Ostrom, E. B. Ebbesen, R. S. Wyer, D. L. Hamilton, & D. E. Carlston (Eds.). *Person Memory: The cognitive basis of social perception* (pp. 155—177). Hillsdale, NJ: Erlbaum.

Hewstone (1994). Revision and change of stereotypic beliefs: In search of the elusive subtyping model. *European Review of Social Psychology, 5*, 70-109.

Hewstone, M., Macrae, C. N., Griffiths, R., & Milne, A. (1994). Cognitive models of stereotype change (5): Measure, developments and consequences of subtyping. *Journal of Experimental Social Psychology, 30*, 505-526.

Ito, T.A., & Cacioppo, J.T. (2001). Affect and attitudes: A social neuroscience approach. In J.P. Forgas (Ed.) *Handbook of affect and social cognition* (pp. 50-74). Mahwah, NJ: Lawrence Erlbaum Associates.

Johnston, L. & Hewstone, M. (1992). Cognitive models of stereotype change: (3) Subtyping and the perceived typicality of disconfirming group members. *Journal of Experimental Social Psychology, 28*, 360—386.

Johnston, L., Hewstone, M., Pendry, L. & Frankish, C. (1994). Cognitive models of stereotype change: (4) Motivational and cognitive influences. *European Journal of Social Psychology*, 24, 237-266.

Kamin, L.J. (1969). Predictability, surprise, attention and conditioning. In B.A. Campbell & R.M. Church (Eds.), *Punishment and aversive behavior*. New York, Appleton Century Crofts.

Kashima, Y., Woolcock, J., & Kashima, E. S. (2000). Group impression as dynamic configurations: The tensor product model of group impression formation and change. *Psychological Review, 107*, 914-942

Kim, H. & Baron, R. S. (1988). Exercise and the illusory correlation: Does arousal heighten stereotypic processing? *Journal of Experimental Social Psychology, 24*, 27—40.

Klauer, K.C., & Meiser, T. (2000). A source-monitoring analysis of illusory correlations. *Personality and Social Psychology Bulletin, 26*, 1074-1093.

Krueger, J., & Clement, R.W. (1994). Memory-based judgments about multiple categories: A revision and extension of Tajfel's Accentuation Theory. *Journal of Personality and Social Psychology, 67*, 35-47.

Kruglanski, A. W., Schwartz, S. M., Maides, S., & Hamel, I. Z. (1978). Covariation, discounting, and augmentation: Towards a clarification of attributional principles. *Journal of Personality, 76,* 176–189.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory**.** *Psychological Review, 103*, 284-308

Lambert, S. B., Klein, O., & Azzi, A. (2002) *Revisiting Tajfel & Wilkes (1963): The joint influence of demand effects and uncertainty*. Paper presented at the Workshop on Social Psychology in Belgium, 18 January 2002. Brussels, Belgium: Royal Flemish Academy of Belgium for Sciences and Arts.

Linville, P.W., & Fischer, G.W. (1993). Exemplar and abstraction models of perceived group variability and stereotypicality. *Social Cognition, 11*, 92-125.

Linville, P.W., Fischer, G.W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer

simulation. *Journal of Personality and Social Psychology, 57*, 165-188.

Mackie, D. M., Hamilton, D. L., Schroth, H. A., Carlisle, C. J., Gersho, B. F., Meneses, L. M., Nedler, B. F., & Reichel, L. D. (1989). The effects of induced mood on expectancy-based illusory correlations. *Journal of Experimental Social Psychology, 25*, 524—544.

Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. *Journal of Personality and Social Psychology, 38*, 231—248.

McClelland, J. M. & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs and exercises*. Cambridge, MA: Bradford.

McClelland, J.L. & Rumelhart, D.E. (1986). A distributed model of human learning and memory. In J. L. McClelland & D.E. Rumelhart (Eds.). *Parallel distributed processing: Explorations in the micro-structure of cognition, Vol. 2* (pp.170—215). Cambridge, MA: MITT Press.

McConnell, A.R., Sherman, S.J., & Hamilton, D.L. (1994). Illusory correlation in the perception of groups: An extension of the distinctiveness-based account. *Journal of Personality and Social Psychology, 67,* 414-429.

McLeod, P., Plunkett, K. & Rolls, E. T. (1998). *Introduction to connectionist modeling of cognitive processes*. Oxford, UK: Oxford University Press.

Meiser, T. & Hewstone, M. (2002). *Cognitive processes in stereotype formation: distinctiveness, statistical reasoning and parallel distributed memory*. Unpublished manuscript.

Messick, D.M. & Mackie, D.M. (1989). Intergroup relations. *Annual review of Psychology, 40,* 45-81.

Mullen, B. & Johnson, C. (1990). Distinctiveness-based illusory correlations and stereotyping: A meta-analytic integration. *British Journal of Social Psychology, 29,* 11-28.

Mullen, B., & Hu, L. (1989). Perceptions of ingroup and outgroup variability: A meta-analytic integration. *Basic and Applied Social Psychology, 10,* 233-252.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115, 39-57.

Nosofsky, R. M., Kruschke, J. K. & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*, 211—233.

Ochsner, K. N., & Lieberman, M. D. (2001). The emergence of social cognitive neuroscience. *American Psychologist, 56*, 717—734.

Park, B. & Hastie, R. (1987). Perception of variability in category development: Instance-versus abstraction-based stereotypes. *Journal of Personality and Social Psychology, 53*, 621—635.

Park, B., & Judd, C.M. (1990). Measures and models of perceived group variability. *Journal of Personality and Social Psychology, 95*, 173-191.

Park, B., Wolsko, C. & Judd, C. M. (2001). Measurement of subtyping in stereotype change. Journal of Experimental Social Psychology, 37, 325—332.

Petty, R. E. & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.). *Advances in experimental social psychology* (Vol. 19, pp. 123—205). San Diego, CA: Academic Press.

Queller, S. & Smith, E. (2002). Subtyping versus bookkeeping in stereotype learning and change: Connectionist simulations and empirical findings. *Journal of Personality and Social Psychology, 82*, 300—313.

Read, S. J. & Marcus-Newhall, A. (1993) Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology, 65*, 429—447.

Read, S.J., & Miller, L.C (Eds.) (1998). *Connectionist and PDP models of Social Reasoning and Social Behavior.* Lawrence Erlbaum.

Read, S.J., & Montoya, J.A. (1999). An autoassociative model of causal reasoning and causal learning: Reply to Van Overwalle's critique of Read and Marcus-Newhall (1993). *Journal of Personality and Social Psychology, 76*, 728—742.

Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.) *Classical conditioning II: Current research and theory* (pp. 64–98). New York: Appleton-Century-Crofts.

Shanks, D. R. (1985a). Continuous monitoring of human contingency judgment across trials. *Memory and Cognition, 13*, 158-167.

Shanks, D. R. (1985b). Forward and backward blocking in human contingency judgment. *Quarterly Journal of Experimental Psychology, 37b*, 1—21.

Shanks, D. R. (1987). Acquisition functions in contingency judgment. *Learning and Motivation, 18*, 147—166.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 433–443.

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology, 48a*, 257—279.

Shanks, D. R., Lopez, F. J., Darby, R. J., Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.) *The psychology of learning and motivation* (Vol. 34, pp. 265—311). New York, NY: Academic Press.

Shavitt, S., Sanbonmatsu, D. M., Smittipatana, S. & Posavac, S. S. (1999). Broadening the conditions for illusory correlation formation: Implications for judging minority groups. *Basic and Applied Social Psychology, 21*, 263—279.

Shultz, T. R. & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review, 103*, 219-240.

Simon, B., & Brown, R.J. (1987). Perceived homogeneity in minority-majority contexts. *Journal of Personality and Social Psychology, 53,* 703-711.

Simon, B., & Hamilton, D.L. (1994). Self-stereotyping and social context: The effects of relative in-group size and in-group status. *Journal of Personality and Social Psychology, 66,* 699-711.

Simon, B., & Pettigrew, T.F. (1990). Social identity and perceived group homogeneity. *European Journal of Social Psychology, 20*, 269-286.

Slusher, M. P., & Anderson, C. A. (1987). When reality monitoring fails: The role of imagination in stereotype maintenance. *Journal of Personality and Social Psychology, 52*, 653—662.

Smith E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review, 99*, 3—21.

Smith, E. R. & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: Simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology, 74*, 21—35.

Smith, E. R. (1996). What to connectionism and social psychology offer each other? *Journal of Personality and Social Psychology, 70*, 893-912.

Smith, E.R. (1991). Illusory correlation in a simulated exemplar-based memory. *Journal of Experimental Social Psychology, 27,* 107-123.

Spears, R. & van der Pligt, J., & Eiser, J.R. (1985). Illusory correlation in the perception of

group attitudes. *Journal of Personality and Social Psychology, 48*, 863—875.

Spears, R., Eiser, J. R. & van der Pligt, J. (1987). Further evidence for expectancy-based illusory correlations. *European Journal of Social Psychology, 17*, 253—258.

Stroessner, S.J., Hamilton, D.L., & Mackie, D.M. (1992). Affect and stereotyping: The effect of induced mood on distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology, 62*, 564-576.

Tajfel, H. (1969). Cognitive aspects of prejudice. *Journal of Social Issues, 25,* 79-97.

Tajfel, H., & Wilkes, A.L. (1963). Classification and quantitative judgment. *British Journal of Psychology, 54,* 101-114.

Thorpe (1994). Localized versus distributed representations. In M. A. Arbib (Ed.) *Handbook of brain theory and neural networks* (pp. 949-952). Cambridge, MA: MIT Press.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving and W. Donaldson (Eds.), *Organization of memory* (pp. 381-403). New York, Academic Press.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

Van Overwalle, F. & Labiouse, C. (2002). *A recurrent connectionist model of person impression formation*. Unpublished manuscript.

Van Overwalle, F. & Van Rooy, D. (1998). An associative approach to causal attribution. In S. J. Read & L. C Miller (Eds.) *Connectionist and PDP models of Social Reasoning and Social Behavior*. Lawrence Erlbaum.

Van Overwalle, F. & Van Rooy, D. (2001). When more observations are better than less: A connectionist account of the acquisition of causal strength. *European Journal of Social Psychology, 31*, 155-175.

Van Overwalle, F. (1998) Causal Explanation as Constraint Satisfaction: A Critique and a Feedforward Connectionist Alternative. *Journal of Personality and Social Psychology, 74*, 312-328.

Van Overwalle, F., & Jordens, K. (2002). An adaptive connectionist model of cognitive dissonance. *Personality and Social Psychology Review, 6*, 204—231.

Van Overwalle, F., Siebler, F. & Labiouse, C. (2002). *A single-system connectionist model of attitude formation and change*. Manuscript submitted for publication.

Van Rooy, D. (2001). *A connectionist learning model of illusory correlation*. Unpublished PhD Dissertation.

Vanhoomissen, T., De Haan, J., Van Overwalle, F. (2001). *Accentuation between groups and memory: A test of a recurrent connectionist model*. Unpublished data.

Wasserman, E. A., Elek, S. M., Chatlosh, D. L., Baker, A. G. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 19*, 174-188

Weber, R., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology, 45*, 961-977.

Welbourne, J. L. (1999). The impact of perceived entativity on inconsistency resolution for groups and individuals. *Journal of Experimental Social Psychology, 35*, 481-508.

Wells, G. L. & Ronis, D. L. (1982). Discounting and augmentation: Is there something special about the number of causes? *Personality and Social Psychology Bulletin, 8*, 566—572.

Whitney, P., Davis, P.A., & Waring, D.A. (1994). Task effects on trait inference: Distinguishing categorization from characterization. *Social Cognition, 12,* 19-35.

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.

Wyer, R.S., & Srull, T.K. (1989). *Memory and Cognition in its social context.* Hillsdale, NJ:
    Erlbaum.

Footnotes

---

[1] Because this is a recurrent network, competition may in principle work also on the (downward) connections from the group to the attribute and episodic nodes. For instance, strong group→attribute connections may hamper de development of episodic→attribute connections. However these latter connections do not play a direct role in our testing procedures. Other sources of competition that involve episodic connections are less likely, because these connections are relatively weak and thus may have little influence.

[2] More detailed information on the input coding for all simulations can be obtained from the second author.

[3] An alternative procedure is based on the assumption that awareness depends on convergence of activation into a stable "attractor" state for the group node (Cleeremans & Jiménez, 2002). The time needed to settle in such an attractor state can be simulated by recording the number of activation updating cycles before an attractor is reached (McLeod, Plunkett & Rolls, 1998). As one might expect, it yielded very similar results. However, in keeping with our general simulation methodology in which multiple cycles are avoided, and because we are not specifically interested in response times but rather more broadly in any measure of episodic memory, we do not report this more elaborated procedure.

[4] Alternatively, in line with an exemplar approach (Linville et al., 1989), one can also activate all group members and read off the resulting attribute activation. This alternative gives very similar results, because the number of members who typify an attribute act as proxy for the strength of the group node with that attribute. Hence, the more members there are, the stronger the group→attribute connection is, resulting in very similar outcomes.

[5] To make sure that the non-linear activation adjustments settled on a stable state, we also conducted simulations with 49 internal cycles (or 50 cycles in total). The results were very similar.

[6] The recurrent model was also able to reproduce attenuation of recency and response dependency in serial position weights as documented by Kashima et al. (2001). More generally, the delta learning algorithm on which the recurrent model was built is well designed to handle most basic forms of category learning (e.g., Gluck & Bower, 1986a; Estes et al., 1989). However, because of space limitations, these issues and simulations are not discussed further.

Table 1:

Overview of the Simulated Group Biases and the Property creating the Bias

Bias	Findings	Property
Group Impression Formation		
1. Size-based Illusory Correlation	• A minority group is seen as more negative despite the fact that the proportion of positive and negative items is identical to a majority group.	*Acquisition*: greater sample size of opposite attributes in majority group
	• Better memory (assignment latencies) for items from a minority category	*Competition*: greater sample size of attributes in majority group discounts episodic weights
2. Expectancy-based Illusory Correlation	More stereotyped judgments despite the lack of an actual correlation	Prior *acquisition* of greater sample size of stereotypical attributes in group caries over to present acquisition
Group Differentiation		
3. Accentuation	• Perceived differences in attributes are pronounced if group membership is correlated with attribute	*Acquisition*: greater sample size of correlated attribute
	• Better memory (of foils) in uncorrelated condition	*Competition*: greater sample size of attributes in correlated condition discounts episodic weights
Changing Group Impressions		
4. Stereotype Change	Group stereotype changes more if stereotype-inconsistent information is dispersed across many members rather than concentrated in a few	*Competition*: greater discounting of inconsistent attribute concentrated in a few members
Group Variability		
5. Group Homogeneity	Outgroup is seen as more homogenous; however ingroup is seen as more homogenous when it is a minority	*Acquisition*: greater sample size of ingroup attribute, unless ingroup is minority

Table 2

Size-based Illusory Correlation: Learning history (based on McConnell et al., 1994, exp. 2)

Trial Frequency	Group		Desirability		Episodic behaviors [a]			
	A	B	+	-	A+	A-	B+	B-
Experimental Phase								
Group A + #8	1	0	1	0	1	0	0	0
Group A - #4	1	0	0	1	0	1	0	0
Group B + #4	0	1	1	0	0	0	1	0
Group B - #2	0	1	0	1	0	0	0	1
Test Phase								
Evaluation of A	1	0	?	-?	0	0	0	0
Evaluation of B	0	1	?	-?	0	0	0	0
Assignment Latencies of A	?	0	0	0	1	0	0	0
	?	0	0	0	0	1	0	0
Assignment Latencies of B	0	?	0	0	0	0	1	0
	0	?	0	0	0	0	0	1

Note. Cell entries denote external node activation. #=Number of times the trial is repeated, +=Desirable, -= Undesirable, ?=resulting test activations (without external activations) averaged across each row. The order of the experimental trials was randomized.

[a] Each type of behavior is shown in a separate column that involves multiple episodic nodes of which a different one is turned on (activation +1) per trial (e.g., in the 1st trial the 1st episodic node is activated, in the 2nd trial the 2nd node and so on).

Table 3

Accentuation: Learning history for correlated and uncorrelated conditions (based on Vanhoomissen et al., 2001)

Trial Frequency	Newspaper		Attitude		Statements in Articles [a, b]			
	A	B	+	-	A+(A+)	A+(B+)	B-(A-)	B- (B-)
				Correlated (Uncorrelated) Condition				
Favorable Articles [a]								
#4	1 (1)	0 (0)	1	0	1	0	0	0
#4	1 (0)	0 (1)	1	0	0	1	0	0
Unfavorable Articles								
#4	0 (1)	1 (0)	0	1	0	0	1	0
#4	0 (0)	1 (1)	0	1	0	0	0	1
				Test Phase				
Attitude Position on Favorable Articles from A and Unfavorable Articles from B								
	.15	0	?	-?	1	0	0	0
	0	.15	?	-?	0	0	0	1
Recognition Memory on Favorable Articles from A and Unfavorable Articles from B								
	?	0	0	1	1	0	0	0
	0	?	1	0	0	0	0	1

Note. Cell entries denote external node activation. #=Number of times the trial is repeated, +=Favorable, -= Unfavorable to a given Attitude Position, ?=resulting test activations (without external activations) averaged across each row. The order of the learning trials was randomized.

[a] Between parentheses are the epsodic node types and activations for the uncorrelated condition. [b] Each type of statement is shown in a separate column that involves multiple episodic nodes of which a different one is turned on (activation +1) per trial

Table 4

Dispersed or Concentrated Stereotype-inconsistent Information: Learning history (Johnston & Hewstone, 1990, exp. 1)

Trial Frequency	Group	Traits		Specific Group Members [a]		
		Consistent	Inconsist.	Confirmer	Mixed	Disconf.
			Pre-experimental Phase			
#10	1	1	0	0	0	0
			Concentrated			
Consistent						
#4	1	1	0	1	0	0
#8	1	1	0	0	1	0
Inconsistent						
#12	1	0	1	0	0	1
			Dispersed			
Consistent						
#6	1	1	0	1	0	0
#4	1	1	0	0	1	0
#2	1	1	0	0	0	1
Inconsistent						
#8	1	0	1	0	1	0
#4	1	0	1	0	0	1
			Test Phase			
Consistent and Inconsistent Trait Ratings of Group						
	1	?	0	0	0	0
	1	0	?	0	0	0
Typicality of Confirmers / Disconfirmers						
	0	?	-?	1	0	0
	0	?	-?	0	0	1

Note. Cell entries denote external node activation. #=Number of times the trial is repeated, Inconsist.= Inconsistent, Disconf.=Disconfirmer, ?=resulting test activations (without external activations) averaged across each row. The order of the learning trials was randomized within each condition.

[a] There were 2 group members who always confirmed the stereotype, 2 who always disconfirmed the stereotype and 4 who showed mixed traits. Each member type is shown in a separate column that involves multiple nodes of which a different one is turned on (activation +1) per one to six trials (see footnote 2).

Table 5

Outgroup and Ingroup Homogeneity: Learning history (based on Simon & Brown, 1987, exp. 1)

Trial Frequency	Group		Attribute		Episodic Behaviors [a]
	Ingroup	Outgroup	High	Low	
			Experimental Phase		
Ingroup (Outgroup) [b]					
#2 (1)	1 (0)	0 (1)	1.0	0	1
#2 (1)	1 (0)	0 (1)	0.8	0	1
#2 (1)	1 (0)	0 (1)	0.5	0.5	1
#2 (1)	1 (0)	0 (1)	0	0.8	1
			Test Phase		
Variability as Range					
	1	0	?	?	0
	0	1	?	?	0

Note. Cell entries denote external node activation. #=Number of times the trial is repeated, ?=resulting test activation (without external activation). The order of the experimental trials was randomized.

[a] Each behavior involves multiple episodic nodes of which a different one is turned on (activation +1) per trial.

[b] Between parentheses are the trial frequencies and activations for the outgroup condition.

Table 6

Fit of the Simulations, including Alternative Encoding and Models

Nr	Bias	Empirical Measure	Original Simulation	Distributed	Feedforward	Non-linear Recurrent
1	Size-based Illusory Correlation	Likeability [a]	1.00	1.00	1.00	1.00
		Assignment RT	.94	.88	.88	< 0 [X]
2	Expectancy-based Illusory Correlation	Frequency	.97	.95	.97	.96
3	Accentuation	Attitude	1.00	1.00	.76 [X]	.96
		Memory (Foils)	.95	.99	.95	< 0 [X]
4	Stereotype Change	Trait	.99	.99	.95	.98
		Typicality	1.00	1.00	< 0 [X]	1.00
5	Group Homogeneity	Range	1.00	1.00	1.00	1.00

Note. Cell entries are correlations between mean simulated values (averaged across randomizations) and empirical data. For the distributed encoding, each concept was represented by 5 nodes and an activation pattern drawn from a Normal distribution with M = activation of the original simulation & SD = .20 (5 such random pattern were run and averaged) and additional noise at each trial drawn from a Normal distribution with M = 0 & SD = .20, and with learning rate = .03 (except .05 for simulation 1). For the Non-linear auto-associative model, the parameters were: E = I = Decay = .15 and internal cycles = 9 (McClelland & Rumelhart, 1988) with learning rate =.20.

[a] The correlations in this row are trivial as only 2 data points are compared and thus necessarily yield only +1 or -1; the correlations in the other rows each involve 4 data points. [X] Predicted pattern was not reproduced.

Figure Captions

Figure 1: Architecture of an auto-associative recurrent network.

Figure 2: Recurrent network for simulations of group bias with two group nodes representing groups A and B, two attribute nodes representing the desirability and undesirability of the behavior and several episodic nodes each representing the unique meaning of one behavioral statement. (Note that not all lateral connections between nodes at the same layer are drawn to avoid cluttering the figure, but all are working during the simulations).

Figure 3: Simulated evaluative strength ($D_{a,b}$=Difference between desirable and undesirable evaluation for group A and B respectively) in an illusory correlation design in which 2 desirable and 1 undesirable behaviors were alternately presented to the network.

Figure 4. Graphical illustration of the mechanisms of competition (with A=group A, B=Group B, D=desirability, F=frequent behaviors, I=infrequent behaviors). Filled nodes are activated at a single trial, empty nodes are not activated. Full lines denote strong connection weights, broken lines denote moderate weights while dotted lines denote weak weights.

Figure 5. Simulation 1: Size-based Illusory Correlation. Observed data from McConnell, Sherman & Hamilton (1994, exp. 2) and simulation results. (Note that in the bottom panel, the scale is reversed so that higher values reflect better memory and, consequently, slower latencies). The human data are from Tables 4 and 5 in "Illusory correlation in the perception of groups: An extension of the distinctiveness-based account" by A.R. McConnell, S.J. Sherman, & D.L. Hamilton, 1994, *Journal of Personality and Social Psychology, 67,* 414—429. Copyright 1994 by the American Psychological Association. Adapted with permission.

Figure 6. Simulation 2: Expectancy-based Illusory Correlation. Observed data from Hamilton & Rose (1980, exp. 1) and simulation results. The human data are from Table 1 in "Illusory correlation and the maintenance of stereotypic beliefs" by D. L. Hamilton & T. L. Rose, 1980, *Journal of Personality and Social Psychology, 39*, 832—845. Copyright 1980 by the American Psychological Association. Adapted with permission.

Figure 7. Simulation 3: Accentuation. Observed data from Vanhoomissen et al. (2001) and simulation results in function of a correlated or uncorrelated condition.

Figure 8. Simulation 4: Dispersed versus Concentrated Stereotype-inconsistent Information. Observed data from Johnston and Hewstone (1992, exp. 1) and simulation results. The human data are from Table 3 in "Cognitive models of stereotype change: (3) Subtyping and the perceived typicality of disconfirming group members" by L. Johnston & M. Hewstone, 1992, *Journal of Experimental Social Psychology, 28*, 360—386. Copyright 1992 by Academic Press. Adapted with permission.

Figure 9. Simulation of Group Variability in function of the high and low extreme of an attribute. The high attribute is shown on the top half, the low attribute on the bottom half. The activation of the low attribute is reverse scored to visualize that variability is measured by range, i.e., the sum of the values obtained for the two opposite attributes.

Figure 10. Simulation 5: Simulation of Ingroup-Outgroup homogeneity in function of (non) minority status of ingroup. Observed data from Simon & Brown (1987) and simulation results. The human data are from the top panel of Figure 1 in "Perceived homogeneity in minority-majority contexts" by B. Simon & R. J. Brown, 1987, *Journal of Personality and Social Psychology, 53,* 703-711. Copyright 1987 by the American Psychological

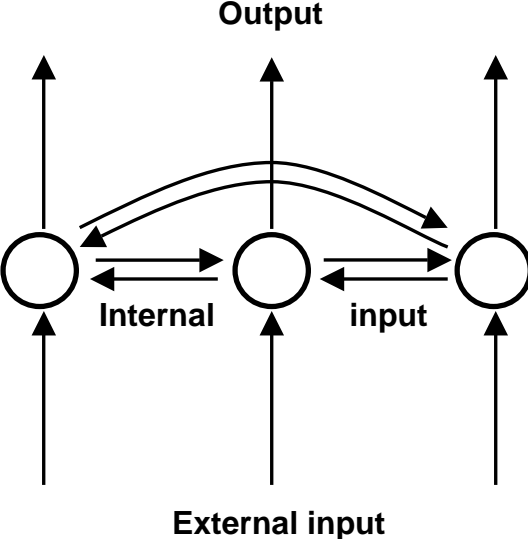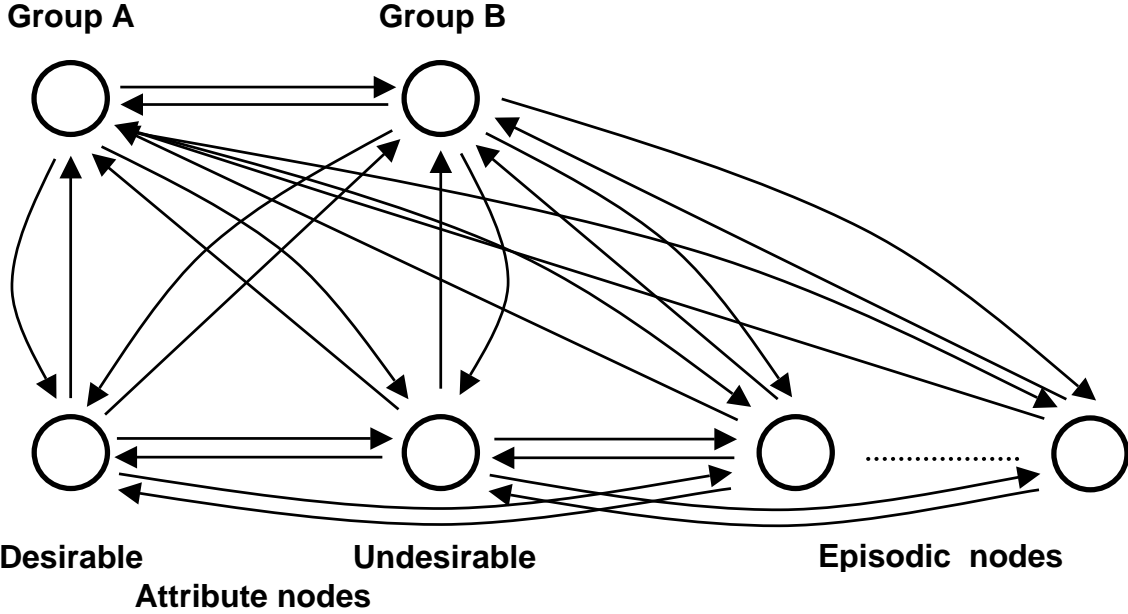Association. Adapted with permission.

Figure 1

Output

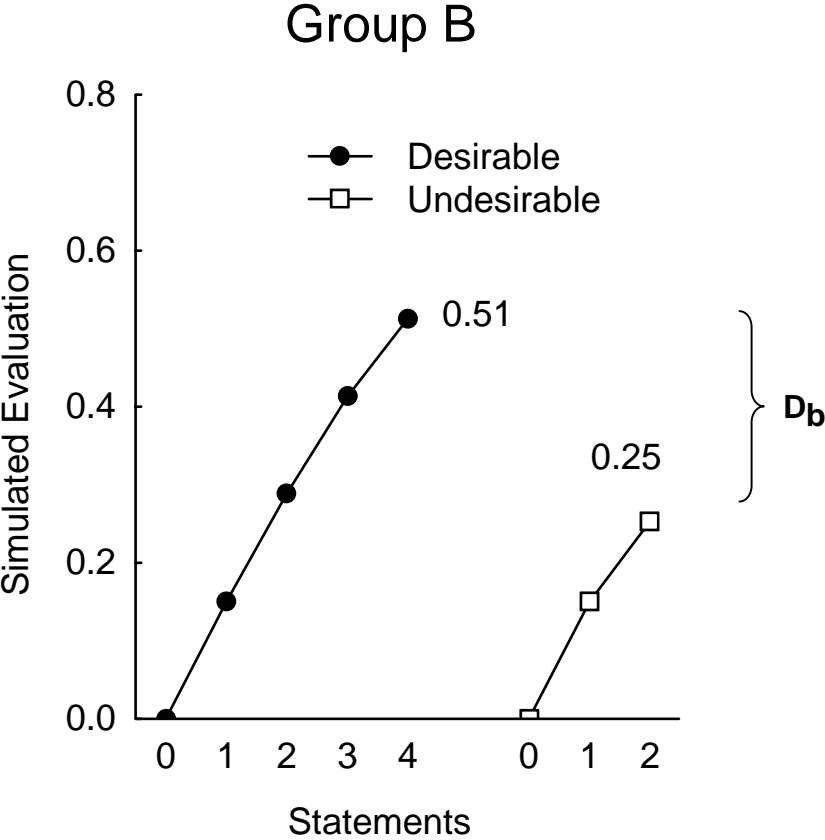**Internal** **input**

External input

Figure 2

Group A   **Group B**

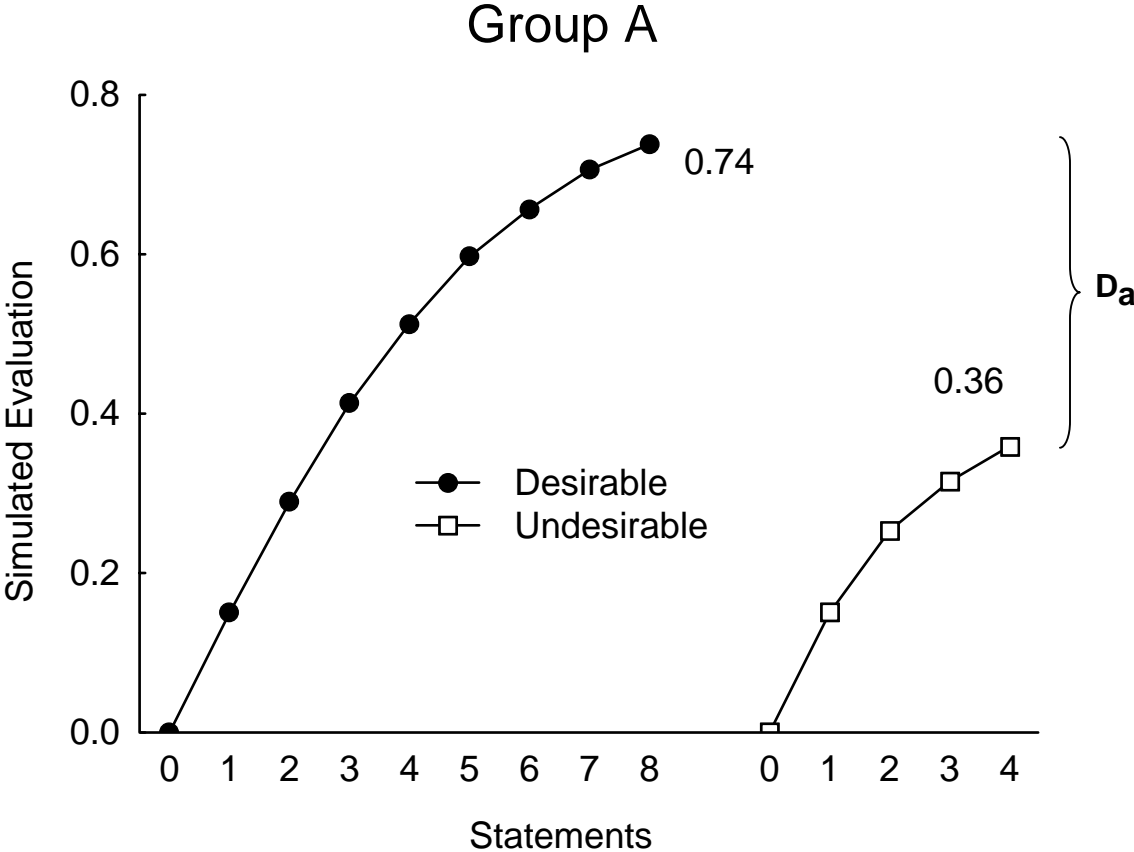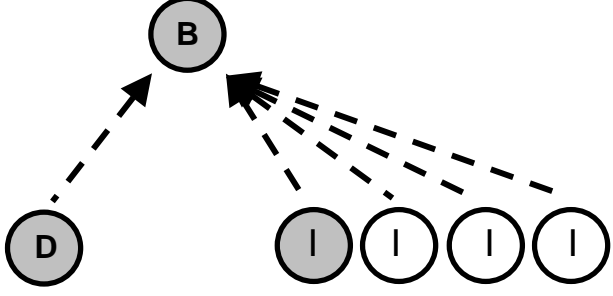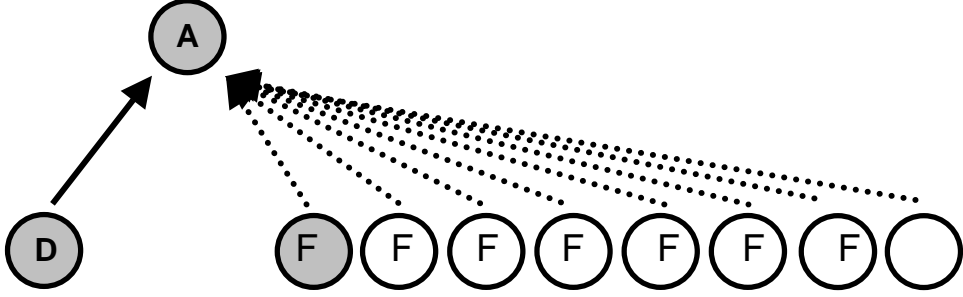**Desirable**   **Undesirable**   **Episodic nodes**
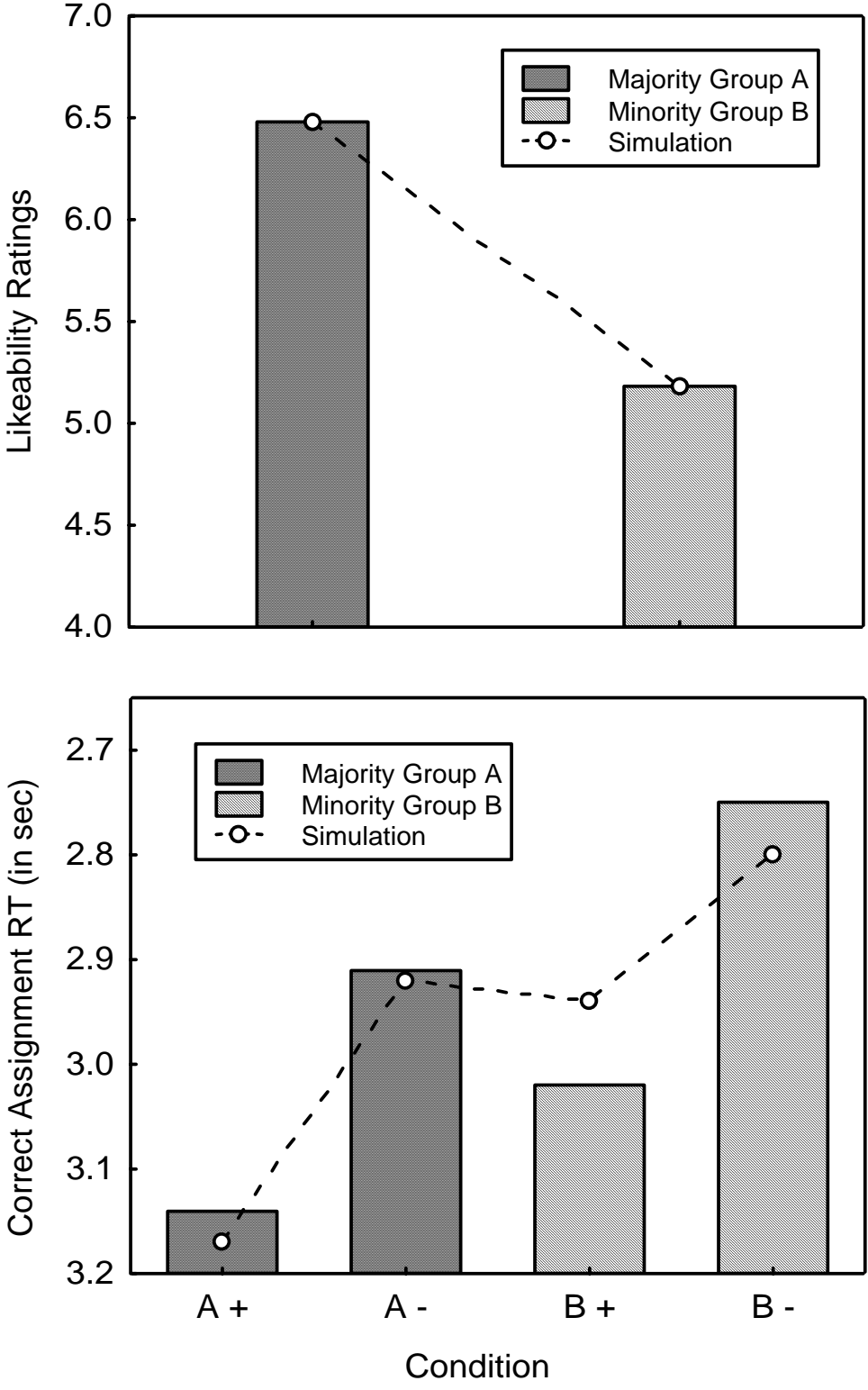
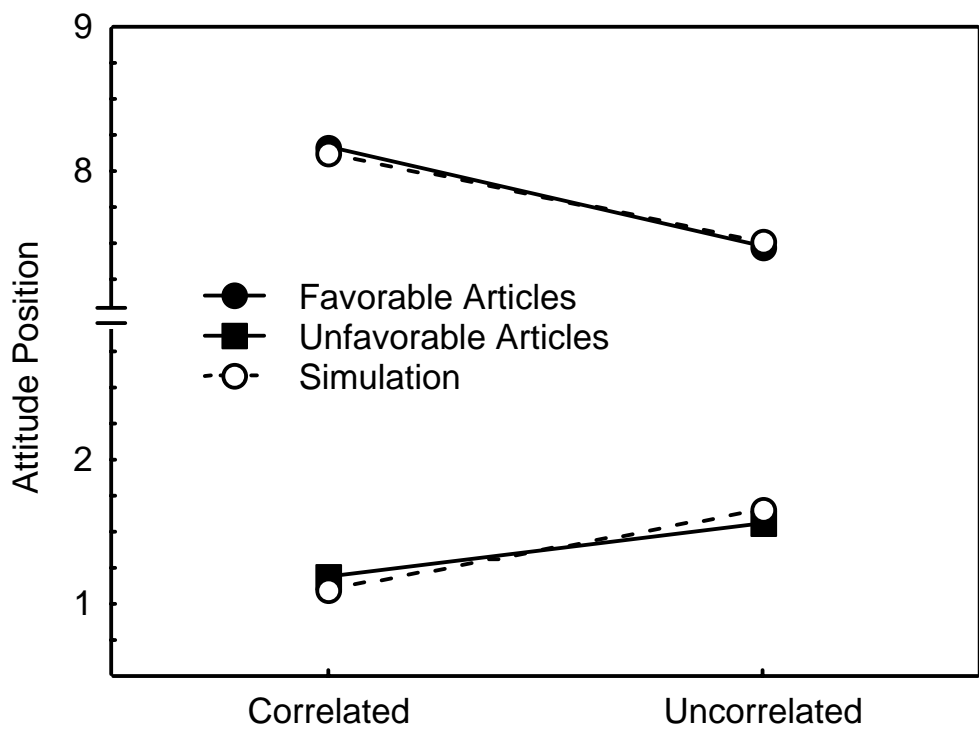**Attribute nodes**

Figure 3

Group A

Group B

Figure 4

Figure 5

Figure 6

Figure 7

Figure 8

Figure 9

Figure 10