Infant Familiarization to Artificial Sentences: Rule-like Behavior Without Explicit Rules and Variables

Shultz, Thomas R.
Bale, Alan C.

Infant Familiarization to Artificial Sentences: Rule-like Behavior Without Explicit Rules and Variables

Thomas R. Shultz (shultz@psych.mcgill.ca)
Department of Psychology; McGill University
Montreal, QC H3A 1B1 Canada

Alan C. Bale (alan_bale@sympatico.ca)
Department of Linguistics; McGill University
Montreal, QC H3A 1G5 Canada

Abstract

A recent study of infant familiarization to artificial sentences claimed to produce data that could only be explained by symbolic rule learning and not by unstructured neural networks. Here we present successful unstructured neural network simulations showing that these data do not uniquely support a rule-based account. In contrast to other neural network simulations, our simulations cover more aspects of the data with fewer assumptions using a more realistic coding scheme based on sonority of phonemes. Our networks show exponential decreases in attention to a repeated sentence pattern, more recovery to novel inconsistent sentences than to novel consistent sentences, some preference reversals, and extrapolation.

One of the most simulated phenomena in developmental psychology is a data set that was claimed to be immune from simulation by unstructured neural networks (Marcus, Vijayan, Bandi Rao, & Vishton, 1999). Although the authors maintained that their results could only be explained by explicit rules and variables, there are now at least eight connectionist simulations of the data, most of which do not use explicit variable binding and none of which use explicit rules. Here we present additional neural simulations of these data, arguing that our model may provide the currently most satisfying account. The paper reviews the relevant infant data, presents various interpretations and models, and then focuses on our current model.

The Infant Data

The relevant experiments familiarized 7-month-old infants to three-word artificial sentences and then tested them on novel sentences that were either consistent or inconsistent with the familiar pattern. The design of these experiments is shown in Table 1. In Experiment 1, infants were familiarized to sentences with either an ABA pattern (e.g., *ni la ni*) or an ABB pattern (e.g., *ta gi gi*). There were 16 of these sentences, constructed by combining four A-category words (*ga*, *li*, *ni*, and *ta*) with four B-category words (*ti*, *na*, *gi*, and *la*). After infants became familiar with a sentence pattern, they were tested with two sentences having novel words that were either consistent or inconsistent with the familiar pattern.

Table 1: Marcus et al. (1999) experiments.

Pattern	Experiments 1 & 2		Experiment 3	
	Cond. 1	Cond. 2	Cond. 1	Cond. 2
Familiarize	ABA	ABB	ABB	AAB
Consistent	ABA	ABB	ABB	AAB
Inconsistent	ABB	ABA	AAB	ABB

When an infant looked at a flashing light to the left or right, a test sentence was played from a speaker situated next to the light. Each test sentence was played until the infant either looked away or 15 s elapsed. Infants attended more to inconsistent novel sentences than to consistent novel sentences, showing that they distinguished the two sentence types.

Experiment 2 was the same except that the words were chosen more carefully so that phoneme sequences were different in the familiarization and test patterns. Experiment 3 used the same words as Experiment 2, but in contrastive syntactic patterns that each duplicated a consecutive word: AAB vs. ABB. The idea was to rule out the possibility that infants might have used the presence or absence of consecutively duplicated words to distinguish sentence types.

In all three experiments, infants attended more to inconsistent than to consistent novel sentences. Our concern is with the best theoretical account of these data. Is the infant cognition based on rules and variables or on connections?

A Rule and Variable Interpretation

Marcus et al. (1999) argued that these grammars could not be learned by the statistical methods common to standard neural networks. They also tried some unsuccessful neural network simulations using Simple Recurrent Networks (SRN). The authors proposed that a only a rule-based model could cover their data. "We propose that a system that could account for our results is one in which infants extract algebra-like rules that represent relationships between placeholders (variables) such as 'the first item X is the same as the third item Y' (p. 79)." They allowed that their data might also be accounted for by structured neural networks that implement explicit rules and variables in a neural style: "The problem is not with neural networks per se but with the kinds of neural networks that are currently popular. These networks eschew explicit representations of variables and relations between variables; in contrast, some less widely discussed neural networks with a very different architecture do incorporate such machinery and thus might form the basis for learning mechanisms that could account for our data (pp. 79-80)."

Psychology of Familiarization

A leading psychological analysis of familiarization assumes that infants build categories for stimuli (Cohen, 1973; Sokolov, 1963). Subsequently, they ignore stimuli that correspond to their categories, and concentrate on stimuli that are relatively novel. These processes are often discussed in terms of recognition memory. If there is substantial recovery to a novel test stimulus, then it is considered novel. But if there is little or no recovery, then the stimulus is considered to be recognized as a member of a familiar category. During familiarization there is typically an exponential decrease in attention.

Familiarization in Neural Networks

Encoder networks that learn to reproduce their inputs on their output units can simulate familiarization and novelty effects in infants (Mareschal & French, 1997). Relations among stimulus features are encoded in hidden unit representations, and accuracy is tested by decoding these hidden unit representations onto output units. Discrepancy between output and input representations is network error. Familiar stimuli produce less error than novel stimuli, which presumably deserve further learning. Such hidden unit representations enable prototypes, generalization, and pattern completion (Hertz, Krogh, & Palmer, 1991).

Other Neural Network Models

There are at least eight alternative computational models of the Marcus et al. (1999) data, all of them connectionist models, presumably attracted by the challenge that ordinary connectionist models would not be able to simulate the data. Most of these models are ordinary unstructured connectionist models without explicit rules and variables. All eight of these models cover the basic finding of the Marcus et al. (1999) experiments, namely noticing the difference between consistent and inconsistent sentences. It is beyond the scope of this brief paper to thoroughly review all of these models, many of which are as yet only sketchily reported. However, we can briefly characterize each model and identify what we believe to be its best virtue and most significant limitation.

Four of the unstructured models use the SRN architecture, construing the network's task to be prediction of the next word in a sentence. Negishi (1999a, b) used an SRN without hidden units, coding each word in analog fashion with place of consonant articulation and vowel height. This is a simple network requiring no unusual hand-wired assumptions or pre-experimental experience. However, Marcus (1999a) claimed that it essentially implemented variables by using continuous values on the input units that are transmitted directly to the outputs, thus arguably disqualifying the model from meeting the challenge that variable binding is required.

Following an argument that Marcus et al.'s (1999) SRNs failed because they lacked normal phonemic experience (Seidenberg & Elman, 1999), Elman (1999) pre-trained an SRN to distinguish whether each word differed or not from the previous word. Each word was coded on 12 binary phonetic features. Although 7-month-olds obviously know something about phonemes and it may be reasonable to include such knowledge in models, it is unlikely that infants receive any target signals about phonemic sameness and difference. More seriously, the network's task in both the pre-training and habituation phases of the simulation was discrimination rather than habituation as it was for the infants.

Christiansen and Curtin (1999) pre-trained an SRN on word segmentation. The network learned to predict the identity and stress of the next phoneme in sentences from information on 11 binary phonological features and the stress and utterance boundaries of individual phonemes. Presented with the Marcus et al. test sentences, the network then showed slightly better prediction of words occurring in inconsistent than those occurring in consistent sentences. Again, the use of prior knowledge seems reasonable. However, it is unclear why the network would perform better on inconsistent sentences, with which it is less familiar, than on consistent sentences whose pattern it has just learned.

Altmann and Dienes (1999) used SRNs with an extra encoding layer between the input and hidden layers. Unlike some models, this one does not require any questionable pre-training and is performing the habituation task. On the negative side, Marcus (1999b) reports that only when somewhat unconventional correlation and distance measures are used can the network discriminate between consistent and inconsistent sentences. It would be more typical to measure error or relative output activation for such networks.

Gasser and Colunga (1999) used a specially-designed network with micro-relation units whose activations correlated with inputs from two different syntactic categories. Hard-wired connections caused similar syllables to be synchronized, producing low activations on the micro-relation units, and dissimilar syllables to be desynchronized, producing high activations on the micro-relation units. No pre-training was necessary, but the hardwiring of connection weights is of questionable psychological validity.

Shastri and Chang (1999; Shastri, 1999) designed a structured connectionist model with explicit variable binding, implemented by temporal synchrony of activations on units representing sequential position and other units representing arbitrary binary word features. The network learned to represent an ABA pattern by firing the first position unit synchronously with the third position unit. This network would seem to generalize well to any novel sentences of three words, regardless of the particular features of the words used. But the network is extensively hand-built, and the critically important feedback signals about the position of words in a sentence are psychologically implausible.

None of the foregoing reports of models include evidence on the course of habituation or provide predictions that could be tested with infants.

Shultz (1999) used an encoder version of the cascade-correlation algorithm with arbitrary analog coding of syllables. With an encoder network, the task is construed as word and sentence recognition. Besides covering the consistency effect, these networks learned the training patterns with an exponential decrease in error and showed occasional reversals of preference that were found with the infants. Because the coding was arbitrary, however, it was not possible to simulate the detailed phonetic differences between Marcus et al.'s (1999) Experiments 1 and 2.

Our Model

Here we present a simulation like that of Shultz (1999), but with phonetically realistic encoding of the input sentences using a continuous sonority scale. A successful result would suggest that such coding could be used by infants in their sentence processing. Sonority is the quality of vowel likeness, and can be defined by perceptual salience (Price, 1980) or by openness of the vocal tract (Selkirk, 1984). The coding scheme is shown in Table 2. The specific numbers are somewhat arbitrary, but their ordering is based on phonological work (Selkirk, 1984; Vroomen, van den Bosch, & de Gelder, 1998).

Table 2: Sonority scale with examples in IPA.

Phoneme category	Examples	Sonority
low vowels	/a/ /æ/	6
mid vowels	/ɛ/ /e/ /o/	5
high vowels	/I/ /i/ /U/ /u/	4
semi-vowels and laterals	/w/ /y/ /l/	-1
nasals	/n/ /m/	-2
voiced fricatives	/z/ /v/	-3
voiceless fricatives	/s/ /f/	-4
voiced stops	/b/ /d/ /g/	-5
voiceless stops	/p/ /t/ /k/	-6

Sonorities range from -6 to 6 in steps of 1, with a gap and change of sign between the consonants and vowels. Each word was coded on two units for the sonority of its consonant and that of its vowel. This is similar to Negishi's (1999b) coding, except that we place consonants and vowels on a single scale, rather than on separate scales. We coded each sentence in the artificial language with six units, two for each one-syllable word. For example, the sentence *ni la ni* was coded as (-2 4 -1 6 -2 4).

Our learning algorithm, cascade-correlation, grows during learning by recruiting new hidden units into the network as required to reduce error (Fahlman & Lebiere, 1990). Recruited hidden units are installed each on a separate layer, receiving input from the inputs and from existing hidden units. The candidate hidden unit that gets recruited is the one whose activations correlate best with current error. After recruiting a hidden unit, the network returns to the phase in which weights feeding the output units are adjusted to reduce error. An encoder option to cascade-correlation (Shultz, 1999) freezes direct input-output connections at 0 to prevent trivial solutions in which weights of about 1 are learned between each input unit and its corresponding output unit.

The cascade-correlation algorithm has been used to simulate many other aspects of cognitive development, including the balance scale (Shultz, Mareschal, & Schmidt, 1994), conservation (Shultz, 1998), seriation (Mareschal & Shultz, 1999), discrimination shift learning (Sirois & Shultz, 1998), pronoun semantics (Oshima-Takane, Takane, & Shultz, 1999), and integration of velocity, time, and distance cues (Buckingham & Shultz, in press).

In these models, network behavior becomes rule-like with learning, but knowledge is clearly not represented in rules and cognitive processing is definitely not accomplished by explicit variable binding and rule firing. Instead, rules are viewed as abstract, epi-phenomenal characterizations of processes occurring at the sub-symbolic level of unit activations and connection weights (Smolensky, 1988).

There are several advantages of implementing rule-like behavior with neural processes, including the acquisition of psychologically realistic non-normative rules, integration of perceptual and cognitive phenomena, natural variation across problems and individuals, and achievement of the right degree of crispness in knowledge representations. In many cases, universally quantified rules are too crisp to model knowledge representations in children.

Neurological justification for generative networks such as cascade-correlation is provided by recent findings on learning-driven neurogenesis and synaptogenesis throughout the lifespan (Quartz & Sejnowski, 1997). Although neurogenesis and neural migration may be too slow to account for learning within the time frame of the typical infant familiarization experiment, there is evidence that synaptogenesis can occur within seconds (Bolshakov, Golan, Kandel, & Siegelbaum, 1997).

Like most models of higher cognition, cascade-correlation is not a model of detailed neural circuits. Instead, it is an abstracted and simplified model that is partly inspired by neural principles. Individual units in cascade-correlation networks may correspond roughly to groups of biological neurons, and connection weights may correspond roughly to neural pathways.

Results

Mean network error on test patterns for the three experiments is shown in Table 3. Main effects of consistency were significant at $p < .0001$. The results show more network error to inconsistent test patterns than to consistent test patterns for each experiment. On the assumption that error represents a need for further cognitive processing, these results capture the infant data.

Table 3: Mean error on test patterns.

Expt.	Patterns	Consistent	Inconsistent
1	ABA v. ABB	8.2	14.5
2	ABA v. ABB	13.1	15.8
3	AAB v. ABB	12.9	15.3

The proportion of networks showing a reversal of the consistency effect was .0667, which is close to the .0625 obtained with infants.

A plot of mean error over epochs for a representative network from the ABB condition of Experiment 1 is shown in Figure 1. The first few epochs are omitted for clarity because error started so high, at around 350. Such plots reveal exponential decreases in error on the training patterns over time, similar to the shape of declining attention in infant familiarization. The epochs at which hidden units are installed are shown with diamond shapes just above the training error. As in most cascade-correlation simulations, error decreases sharply after a hidden unit is recruited. After training, error is higher on inconsistent test patterns than on consistent test patterns.

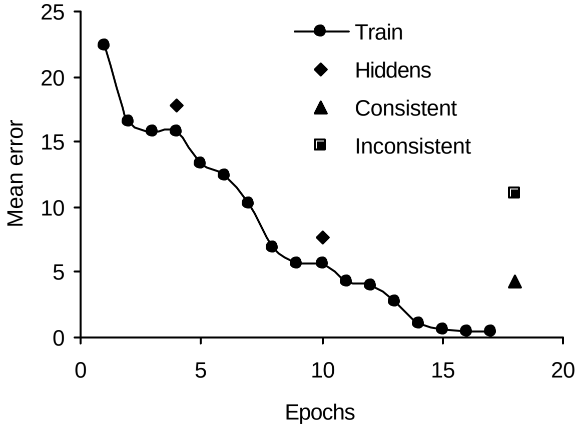

Figure 1: Error reduction in one network.

Generalization tests show that the consistency effect actually grows larger with increasing distance from the training set, a prediction quite different than universally quantified rules would make.

Network analysis revealed that hidden units used sonority sums of consonant and vowel to represent sonority variation first in the duplicated-word category and second in the single-word category. Networks decoded this hidden unit representation with virtually duplicate weights to outputs representing the duplicate-word category.

Discussion

Like other neural models, our model easily captures the consistency effect. In contrast to alternate models of these data, ours has several features to recommend it. Our model does not require extensive pre-experiment experience (Christiansen & Curtin, 1999; Elman, 1999), extensive hand-wiring of networks (Gasser & Colunga, 1999; Shastri & Chang, 1999), external feedback signals not available in the stimuli (Elman, 1999; Shastri & Chang, 1999), unusual interpretation of outputs (Altmann & Dienes, 1999), or explicit variable binding (Shastri & Chang, 1999). On grounds of theoretical parsimony, the more unsupported assumptions that a model requires the less plausible it becomes.

Unlike some alternate models (Shastri & Chang, 1999; Shultz, 1999), our model uses a realistic coding of the stimuli. Like Negishi (1999b), we used an analog coding of inputs based on the manner in which the phonemes are produced. But our representation scheme is a bit more compact and uniform because we use a single sonority scale for both consonants and vowels, whereas he used two separate scales, one for place of consonant articulation and another for vowel height. Moreover, our use of hidden units with non-linear transfer functions ensures that any possible variable binding at the input level is lost as activation is propagated forward through the hidden layers.

Our model is the only one so far to capture the other feature of the Marcus et al. (1999) infant data, the occasional reversal of preference for novel patterns. It is unclear how easily other models might be able to capture these reversals, but there are hints that it might be difficult for some models. Elman's (1999) model, for example, had such a strong consistency effect that reversals of preference would be unlikely: mean activation to ABB sentences was 123 times higher than to ABA sentences. Likewise, the Shastri and Chang (1999) model learns a very strong representation of serial position. The correlation between weights to position nodes were .9993 for positions 1 and 3 in networks habituated to ABA sentences, and .9998 for positions 2 and 3 in networks habituated to ABB sentences. This rather crisp representation produced 3.4 times more error to inconsistent than to consistent sentences in the ABA condition of Experiment 1, which would seem to preclude reversals.

Although it is not known why infants show occasional reversals, our simulations show that they can be a natural part learning. With limited exposure, as in both the psychological experiments and our simulations, exceptions naturally occur. This is a parsimonious explanation of reversals because it does not require assumptions of any extraneous processes.

In summary, our model might be currently preferred because it covers more of the infant data, with less pre-experimental experience, less network design, and more realistic stimulus coding than alternate models. It also uses a general learning algorithm that has been applied successfully to several other phenomena in cognitive development.

With so many successful neural models of the consistency effect, there is no question that ordinary, unstructured neural networks can cover these data. The modeling shows that some of the functionality of symbolic rules and variable binding can be constructed from sub-symbolic processes without having to be explicitly built in. The time is now ripe to generate and test predictions from these alternate models.

Acknowledgments

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. We thank David Buckingham, Jacques Katz, Yuriko Oshima-Takane, Sylvain Sirois, and Yoshio Takane for comments.

References

Altmann, G. T. M., & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science, 284*, 875.

Bolshakov, V. Y., Golan, H., Kandel, E. R., & Siegelbaum, S. A. (1997). Recruitment of new sites of synaptic transmission during the cAMP-dependent late phase of LTP at CA3-CA1 synapses in the hippocampus. *Neuron, 19*, 635–651.

Buckingham, D., & Shultz, T. R. (in press). The developmental course of distance, time, and velocity concepts: A generative connectionist model. *Journal of Cognition and Development*.

Christiansen, M. H., & Curtin, S. L. (1999). The power of statistical learning: No need for algebraic rules. *Proceedings of the Twenty-first Annual conference of the Cognitive Science Society* (pp. 114-119). Mahwah, NJ: Erlbaum.

Cohen, L. B. (1973). A two-process model of infant visual attention. *Merrill-Palmer Quarterly, 19*, 157-180.

Elman, J. L. (1999). Generalization, rules, and neural networks: A simulation of Marcus et al. www.crl.ucsd.edu/~elman/Papers/MVRVsim.html

Fahlman, S. E., & Lebiere, C. (1990). The Cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems 2* (pp. 524-532). Los Altos, CA: Morgan Kaufmann.

Gasser, M., & Colunga, E. (1999). Babies, variables, and connectionist networks. *Proceedings of the Twenty-first Annual conference of the Cognitive Science Society* (p. 794). Mahwah, NJ: Erlbaum.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Reading, MA: Addison Wesley.

Marcus, G. F. (1999a). Do infants learn grammar with algebra or statistics? *Science, 284*, 433.

Marcus, G. F. (1999b). Response: Rule learning by seven-month-old infants and neural networks. *Science, 284*, 875.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283*, 77-80.

Mareschal, D. & French, R. M. (1997). A connectionist account of interference effects in early infant memory and categorization. *Proceedings of the 19th annual conference of the Cognitive Science Society* (pp. 484-489). Mahwah, NJ: LEA.

Mareschal, D., & Shultz, T. R. (1999). Development of children's seriation: A connectionist approach. *Connection Science, 11*, 149-186.

Negishi, M. (1999a). Do infants learn grammar with algebra or statistics? *Science, 284*, 433.

Negishi, M. (1999b). Rule learning by seven-month-old infants and by a simple-recurrent-network. www.cns-web.bu.edu/pub/mnx/sci.html

Oshima-Takane, Y., Takane, Y., & Shultz, T. R. (1999). The learning of first and second pronouns in English: Network models and analysis. Journal of Child Language, 26, 545-575.

Price, P.J. (1980). Sonority and syllabicity: Acoustic correlates of perception. *Phonetica, 37*, 327-343.

Quartz, S. R, & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioural and Brain Sciences, 20*, 537-596.

Seidenberg, M. S., & Elman, J. L. (1999). Do infants learn grammar with algebra or statistics? *Science, 284*, 433.

Selkirk, E.O. (1984). On the major class features and syllable theory. In M. Aronoff & R.T. Oehrle (Eds). *Language sound structure* (pp. 107-136). Cambridge MA: MIT Press.

Shastri, L. (1999). Infants learning algebraic rules. *Science, 285*, 1673.

Shastri, L., & Chang, S. (1999). A spatiotemporal connectionist model of algebraic rule-learning. TR-99-011. International Computer Science Institute, Berkeley, CA. www.icsi.berkeley.edu/~shastri/babytalk

Shultz, T. R. (1998). A computational analysis of conservation. *Developmental Science, 1*, 103-126.

Shultz, T. R. (1999). Rule learning by habituation can be simulated in neural networks. *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society* (pp. 665-670). Mahwah, NJ: Erlbaum.

Shultz, T. R., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning, 16*, 57-86.

Shultz, T. R., Oshima-Takane, Y., & Takane, Y. (1995). Analysis of unstandardized contributions in cross connected networks. In D. Touretzky, G. Tesauro, & T. K. Leen, (Eds). *Advances in Neural Information Processing Systems 7* (pp. 601-608). Cambridge, MA: MIT Press.

Sirois, S., & Shultz, T. R. (1998). Neural network modeling of developmental effects in discrimination shifts. *Journal of Experimental Child Psychology, 71*, 235-274.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1-74.

Sokolov, E. N. (1963). *Perception and the conditioned reflex*. Hillsdale, NJ: Erlbaum.

Vroomen, J., van den Bosch, A., & de Gelder, B. (1998). A connectionist model for bootstrap learning of syllabic structure. *Language and Cognitive Processes, 13*, 193-220.

Simulation of Self-affirmation Phenomena in Cognitive Dissonance

Thomas R. Shultz (shultz@psych.mcgill.ca)
Department of Psychology; McGill University
Montreal, QC H3C 1B1 Canada

Mark R. Lepper (lepper@psych.stanford.edu)
Department of Psychology; Stanford University
Stanford, CA 94305-2130 USA

Abstract

The consonance constraint-satisfaction model, which has simulated the major paradigms of classical cognitive dissonance theory, is here extended to deal with more contemporary findings concerning self-affirmation phenomena in dissonance reduction. The key addition to the model, which has also figured in recent simulations of arousal phenomena, is to lessen activity level within the neural network model in self-affirmation conditions. These and other simulations continue to show that dissonance phenomena can be explained in terms of constraint satisfaction.

Introduction

One of the fundamentally important theories in social psychology is cognitive dissonance theory, which has generated a literature of more than 1000 studies over the past 40 years (Festinger, 1957; Thibodeau & Aronson, 1992). We have recently modeled a number of the central dissonance phenomena using constraint-satisfaction neural networks (Shultz & Lepper, 1996, 1998a&b, 1999a&b). Our so-called consonance model covered insufficient justification, free choice, arousal, and some self-concept phenomena. The model also predicted new free-choice effects that were subsequently confirmed by further psychological experimentation (Shultz, Léveillé, & Lepper, 1999). In this paper, we report on an extension of the model to deal with a prominent self-concept effect in dissonance called self-affirmation.

Dissonance is hypothesized to occur when behavior is inconsistent with self-concept (Steele, 1988; Thibodeau & Aronson, 1992). Because most people have a positive self-concept, behaviors such as lying or trying to persuade others of a position that one does not agree with arouse dissonance and lead to attitude change that reduces the dissonance. However, if important aspects of the self-concept have been recently affirmed, even aspects irrelevant to an experimentally induced inconsistency, there may be no need to reduce dissonance via attitude change. Steele (1988) presented experiments in which fairly subtle self-affirmation manipulations eliminated dissonance effects. Some of these experiments concern insufficient justification via forced compliance, and others deal with free choice. We return to these experiments after reviewing the consonance model used in the simulations.

The Consonance Model

The consonance model holds that dissonance reduction is a constraint satisfaction problem. The motivation to reduce dissonance stems from the various soft constraints on the beliefs and attitudes that an individual holds. A consonance network corresponds to a person's representation of the situation created in the conditions of a dissonance experiment. Activations of network units represent the direction and strength of a person's cognitions. Weights between cognitions represent psychological implications. These unit activations and weights may vary across the different conditions of a single experiment.

Consonance is the degree to which similarly evaluated units are linked by excitatory weights and oppositely valued units are linked by inhibitory weights. More formally, consonance in a network is defined by

$$consonance = \sum_i \sum_j w_{ij} a_i a_j$$

where w_{ij} is the weight between units i and j, a_i is the activation of the receiving unit i, and a_j is the activation of the sending unit j.

Activation spreads over time cycles by two update rules:

$$a_i(t+1) = a_i(t) + net_i(ceiling - a_i(t)) \text{ when } net_i \geq 0$$

$$a_i(t+1) = a_i(t) + net_i(a_i(t) - floor) \text{ when } net_i < 0$$

where $a_i(t+1)$ is the activation of unit i at time $t+1$, $a_i(t)$ is the activation of unit i at time t, *ceiling* is the maximum activation, *floor* is the minimum activation, and net_i is the net input to unit i, defined as:

$$net_i = resist_i \sum_j w_{ij} a_j$$

where $resist_i$ refers to the resistance of receiving unit i to having its activation changed.

At each time cycle, n units (normally the number of units in the network) are randomly selected and updated. The update rules ensure that consonance increases or stays the same across cycles. Consonance increases because positive net inputs drive unit activations toward the ceiling and negative net inputs drive them toward the floor. Consonance increases until units reach extreme values or net inputs fall to 0. When consonance reaches asymptote, updating stops.

Consonance networks are hand-built to implement particular dissonance experiments using a set of five principles that map dissonance theory to the consonance model:

1. A cognition is implemented by the net activation of a pair of negatively connected units, one of which represents the positive aspect and the other the negative aspect of the cognition.
2. Cognitions are connected to each other based on their causal implications.
3. Dissonance is the negative of consonance divided by the number of nonzero inter-cognition relations.
4. Networks settle into more stable, less dissonant states as unit activations are updated.
5. Unit activations, but not connection weights, are allowed to change, and some cognitions are more resistant to change than others. In particular, beliefs, behaviors, and justifications are more resistant to change than are evaluations and attitudes.

Additional details about the consonance model and its assumptions are available in our previous papers (Shultz & Lepper, 1996, 1998a).

Forced Compliance

Forced compliance is the most popular dissonance technique within the most prominent dissonance paradigm of insufficient justification. Insufficient justification concerns cases in which a person does something inconsistent with his or her attitudes without much justification. The less the justification, the more cognitive dissonance is created.

In a forced-compliance experiment (Steele, 1988, p. 272), college students were selected for their strong opposition to an increase in tuition fees. They were then persuaded to write essays supporting a substantial tuition increase. In one condition, they were given a choice about whether to write the essay; in another condition, they were given very little choice about whether to write the essay. When a person freely agrees to argue against personal beliefs, this creates dissonance, which can be reduced by changing attitudes in the direction of the argument. There should be little or no dissonance when one is pressured to make such arguments.

Before measuring post-experimental attitudes, some participants were first asked to complete the political sub-scale of the Allport-Vernon Study of Values. One-half of them had been previously assessed as having a strong economic-political value orientation, whereas the others did not have this value orientation. Completing the political value scale was supposed to affirm a valued self-concept only for those students with a strong economic-political value orientation.

As shown by the solid line in Figure 1, there was the familiar dissonance effect of more attitude change under high choice than under low choice. Moreover, as predicted, self-affirmation eliminated attitude change, even under high choice conditions. Two other experiments with minor variations yielded similar results (Steele, 1988).

Method

Network specifications for the three conditions are shown in Table 1. There are two relevant cognitions, attitude and essay, and relations between them. As in our previous simulations, each cognition is implemented with a pair of negatively related units, one to represent the positive aspect of the cognition and the other to represent the negative aspect. Net activation for a cognition is computed as activation on the positive unit minus activation on the negative unit. Positive relations between cognitions are implemented by positive weights between their positive units and between their negative units, and negative weights between the positive unit of one cognition and the negative unit of the other cognition. All weights are bi-directional.

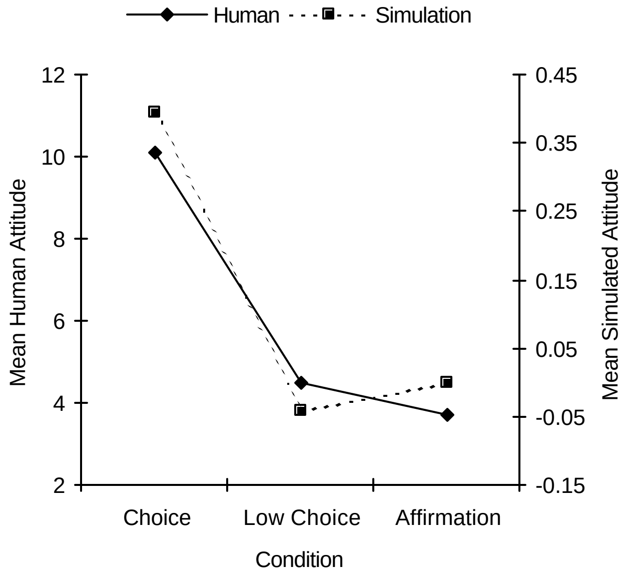

Figure 1: Mean attitude following forced compliance.

All weights and initial unit activations are assigned either high (0.5) or low (0.1) values, according to the five mapping principles described earlier and the descriptions of the experiments being modeled. The floor parameter is 0; the ceiling parameter for positive units is set to 1, and that for negative units is set to 0.5. A *cap* parameter is set to -0.5. This corresponds to the value of the weight between each unit and itself and it prevents activations from growing to ceiling. The *resist* parameter is set to 0.5 for low resistance, and 0.01 for high resistance. These parameter settings are standard across all our dissonance simulations, and some justification for them is provided in our longer papers, (Shultz & Lepper, 1996, 1998a, 1999a).

Table 1: Network specifications for forced compliance.

Condition	Attitude	Essay	Relation
Choice	-0.5	0.5	0.5
Low Choice	-0.5	0.5	0.1
Affirmation	-0.25	0.25	0.25

In this experiment, there is a positive relation between attitude and essay because the more positive one's attitude toward tuition increases, the more likely one would be to agree to write an essay in favor of tuition increases. This relation is high in the choice condition and low in the low-choice condition. Initially, attitude is given a high negative value to reflect students' initial attitudes; and essay is given a high positive value because the essay was indeed written by all students. An activity-level scalar of 0.5 (the same value used in our other simulations of arousal and self concept) reduces initial activations and weights in the self-affirmation condition, relative to the no-affirmation conditions. The theoretical justification for using a scalar in this way is that self-affirmation is hypothesized to reduce the importance of a dissonant situation (Steele, 1988, p. 292).

All initial unit activations and weights are randomized for each network by adding or subtracting a random proportion of their initial amounts. The three proportion ranges in which additions or subtractions are randomly selected under a uniform distribution are .1, .5, and 1. This increases psychological realism because not everyone can be expected to share the same parameter values. It also allows a test of robustness of the model. Twenty networks were run in each condition at these three different levels of parameter randomization. Networks were run for 30 cycles, which was sufficient to approach asymptotic activation levels.

Results

Mean attitude toward the view espoused in the essay is presented, in the dashed line in Figure 1, for networks at the .5 level of parameter randomization. As with Steele's (1988) subjects, attitudes are more positive under choice than under the other two conditions. An ANOVA with condition as the single factor revealed significant main effects of condition, $F(2, 57) = 67$, $p < .001$. A contrast F with weights of +2 for choice, -1 for low choice, and -1 for self-affirmation is significant $F(1, 57) = 135$, $p < .001$, with no significant residual, $F(1, 57) < 1$. Proportion of total variance accounted for by this F is .99.

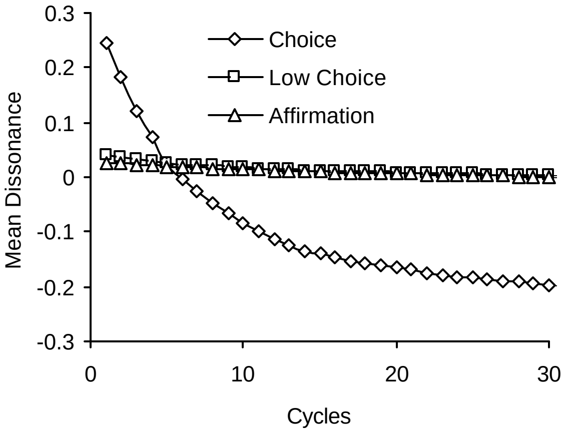

Figure 2: Mean dissonance following forced-compliance.

Mean dissonance scores over time cycles, for networks run at .5 parameter randomization for the three conditions, are shown in Figure 2. Dissonance starts high in the choice condition and is greatly reduced over time. In contrast, there is minimal dissonance in the other two conditions and very little dissonance reduction. Similar results were obtained at parameter randomization levels of .1 and 1.

Discussion

The consonance networks provide a good fit to the attitude change data reported by Steele (1988). There is considerable attitude change in the choice condition, but very little in the low-choice and self-affirmation conditions. There is also a close correspondence between amount of attitude change and plots of dissonance reduction in that the condition with sharp dissonance reduction is also the one with the most attitude change. Examination of dissonance plots is a bonus of computer simulations -- there is no known way to measure dissonance directly in humans. Such plots of simulated dissonance can help to understand the more indirect attitude-change effects that occur as a way of reducing dissonance.

Free Choice

Steele (1988, p. 276) also presents a free-choice experiment that shows self-affirmation effects. Participants rated and ranked 10 music albums and were then given a choice to keep either their fifth- or sixth-ranked album. Choosing between qualitatively distinct objects creates dissonance because the chosen object is less than perfect and the rejected object has some desirable features that are forgone when an irreversible choice is made. The dissonance arising from a free choice is typically reduced by increasing evaluation of the chosen object and decreasing evaluation of the rejected object (Brehm, 1956; Shultz et al., 1999).

In Steele's experiment, one-half of the participants had been previously selected for having a strong scientific-value orientation and for indicating that a lab coat symbolized these values. The others did not share these values. One-half of the participants in each of these groups were asked to wear a lab coat for the rest of the experiment, during which they rated the albums again, after making their choices.

Post-decisional spread of alternatives was measured by adding the increase in the value of the chosen item and the decrease in the value of the rejected item. There were three control conditions, one with participants not having a science orientation and not wearing a lab coat, another with participants not having a science orientation but wearing a lab coat, and a third with participants having a science orientation but not wearing a lab coat. There were identical dissonance effects in all three control conditions, but not for the self-affirmed, scientifically-oriented students wearing a lab coat. Mean spread of alternatives was higher in the control conditions than in the self-affirmation condition, as shown by the solid line in Figure 3. Once again, apparently irrelevant self-affirmation precluded dissonance reduction.

Method

Network specifications for these two groups of conditions are shown in Tables 2 and 3. There are three cognitions: a decision and evaluations of the chosen and the rejected objects. Because the decision is public and irreversible, it has high resistance and high initial activation; the two evaluations have low resistance. Initial evaluation of the chosen object is somewhat higher than that for the rejected object because people generally choose items that they rate higher.

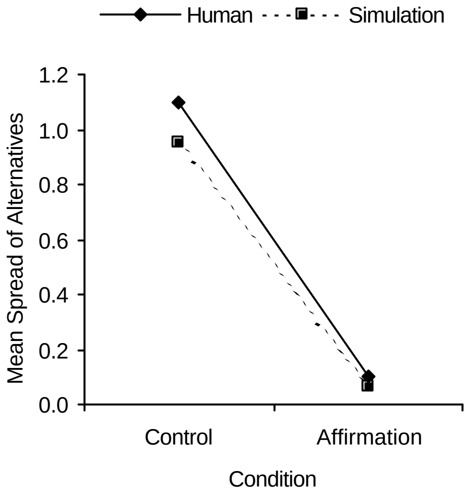

Figure 3: Mean spread of alternatives following free choice.

The relation between the decision and the chosen object is positive because the better-liked object is chosen. The two objects are negatively related because they compete for an exclusive choice. Both relations have high values in the control condition. To implement self-affirmation, initial activations and weights are scaled by .5. Networks in each condition were run for 40 cycles, which was sufficient for saturation. As is customary in our simulations, all weights and initial unit activations were randomized at up to .1, .5, or 1 of the values shown in Tables 2 and 3. Other parameter settings are also the same as in our other dissonance simulations.

Table 2: Initial net activations for free choice.

	Condition	
Cognition	Control	Affirmation
Chosen	.30	.15
Rejected	.20	.10
Decision	.50	.25

Results

Spread between evaluations of the two choices was computed as in Steele (1988). Change in evaluation of each object is the difference between initial evaluation and evaluation after 40 cycles. Spreading of alternatives is the sum of the increase in evaluation of the chosen alternative and the decrease in evaluation of the rejected alternative. Mean spreading of the alternatives is plotted, on the dashed line in Figure 3, at the .5 level of parameter randomization. There is a larger spread of the alternatives in the control than in the self-affirmation condition, $F(1, 38) = 76, p < .001$.

Mean dissonance scores across time cycles in networks at .5 parameter randomization are shown in Figure 4 for the two conditions. Although dissonance starts low in both conditions, it drops only in the control condition. Similar results were found at parameter randomizations of .1 and 1.

Table 3: Relations between cognitions for free choice.

Relation of	Condition	
chosen to	Control	Affirmation
Decision	.50	.25
Rejected	-.50	-.25

Discussion

Consonance networks yield greater separation of alternatives in the control than in the self-affirmation condition, as found with human participants (Steele, 1988). Dissonance reduction is also greater in the control than in the self-affirmation condition, consistent with the idea that attitude change is driven by dissonance reduction.

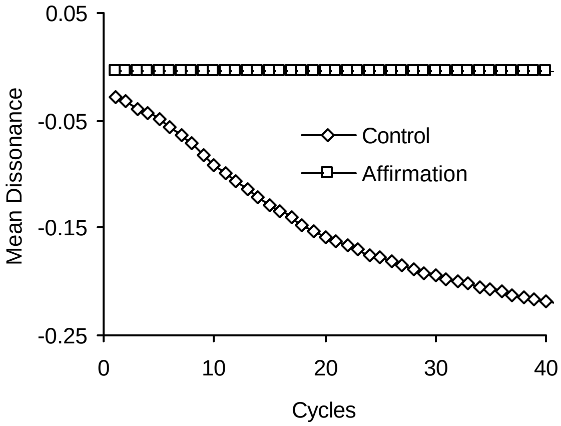

Figure 4: Mean dissonance following free choice.

General Discussion

These simulations extend the consonance model to rather subtle aspects of dissonance reduction involving the self-concept, using the same conventions, mapping principles, and default parameter values as in previous simulations. In all of these cases, dissonance arises when constraints between simultaneously held cognitions are unsatisfied. Dissonance is reduced as the constraints are satisfied, typically by changing evaluations of entities in the situation defined by the dissonance experiment. The self-affirmation phenomena considered here had not previously been simulated and were not generally seen as being closely related to other contemporary dissonance phenomena on emotional arousal. As in earlier simulations, the consonance model is here shown to be robust against parameter variation, as revealed by the fact that even a high degree of parameter randomization does not affect the pattern of overall results.

A key, unifying concept in simulating contemporary dissonance phenomena in self-concept and arousal is that of activity level. An activity scalar adjusts the overall level of activation in networks that represent dissonant situations. In the present simulations, the activity-level scalar operates much like a tranquilizing drug in arousal simulations (Shultz & Lepper, 1999b), by decreasing activation of the representation of the dissonant situation.

Self-affirmation manipulations are thus hypothesized to decrease the relative importance of being in a dissonant situation. When you feel good about yourself, being in a dissonant situation is not nearly so bothersome, and you become immune to the effects of dissonance reduction. This reveals a somewhat unexpected theoretical communality between arousal and self-concept effects.

This analysis is consistent with recent results on *trivialization* as a mode of dissonance reduction (Simon, Greenberg, & Brehm, 1995). Merely making salient to participants asked to write counter-attitudinal essays the contrast between issues they believe to be of great consequence and the less important topic of their own essays reduces attitude change in the direction of the position advocated.

At the level of the brain or an artificial neural network, the key theoretical notion is that of activity level. Dissonance effects are enhanced by increases in activity level and dampened by decreases in activity level. There are a variety of ways to modulate activity level, including general manipulations such as drugs (Cooper, Zanna, & Taves, 1978) and specific manipulations such as attention to particular cognitions (Read & Miller, 1998a). Consequently, activity level has the potential to unify theoretical understanding of several apparently different dissonance phenomena.

The general success of the consonance model enables a theoretical reinterpretation of dissonance that stresses commonalties with other psychological phenomena that result from constraint satisfaction. Phenomena such as analogical reasoning, person perception, schema completion, attitude change, and dissonance reduction can all be understood in terms of the dynamics of constraint satisfaction (Holyoak & Thagard, 1989; Read & Miller, 1998a, b; Rumelhart, Smolensky, McClelland, & Hinton, 1986; Spellman & Holyoak, 1992; Spellman, Ullman, & Holyoak, 1993; Thagard, 1989).

Acknowledgments

This research was supported by a grant to the first author from the Social Sciences and Humanities Research Council of Canada and by grant MH-44321 to the second author from the U.S. National Institute of Mental Health.

References

Brehm, J. W. (1956). Post-decision changes in the desirability of choice alternatives. *Journal of Abnormal and Social Psychology, 52*, 384-389.

Cooper, J., Zanna, M. P., & Taves, P. A. (1978). Arousal as a necessary condition for attitude change following forced compliance. *Journal of Personality and Social Psychology, 36*, 1101-1106.

Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*, 295-355.

Read, S. J., & Miller, L. C. (1998a). On the dynamic construction of meaning: An interactive activation and competition model of social perception. In S. J. Read & L. C. Miller (Eds.). *Connectionist models of social reasoning and social behavior* (pp. 27-68). Hillsdale, NJ: Erlbaum.

Read, S. J., & Miller, L. C. (Eds.). (1998b). *Connectionist models of social reasoning and social behavior.* Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. (1986). Schemata and sequential thought processes in PDP models. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 7-57). Cambridge, MA: MIT Press.

Shultz, T. R., & Lepper, M. R. (1996). Constraint satisfaction modeling of cognitive dissonance phenomena. *Psychological Review, 103*, 219-240.

Shultz, T. R., & Lepper, M. R. (1998a). The consonance model of dissonance reduction. In S. J. Read & L. C. Miller (Eds.), *Connectionist models of social reasoning and social behavior* (pp. 211-244). Hillsdale, NJ: Erlbaum.

Shultz, T. R., & Lepper, M. R. (1998b). A constraint-satisfaction model of Machiavellianism effects in cognitive dissonance. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 957-962). Hillsdale, NJ: Erlbaum.

Shultz, T. R., & Lepper, M. R. (1999a). Computer simulation of cognitive dissonance reduction. In E. Harmon-Jones & Mills, J. (Eds.), *Cognitive dissonance: Progress on a pivotal theory in social psychology* (pp. 235-265). Washington, DC: American Psychological Association.

Shultz, T. R., & Lepper, M. R. (1999b). Consonance network simulations of arousal phenomena in cognitive dissonance. *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society* (pp. 659-664). Hillsdale, NJ: Erlbaum.

Shultz, T. R., Léveillé, E., & Lepper, M. R. (1999). Free choice and cognitive dissonance revisited: Choosing "lesser evils" vs. "greater goods." *Personality and Social Psychology Bulletin, 25*, 40-48.

Simon, L., Greenberg, J., & Brehm, J. (1995). Trivialization: The forgotten mode of dissonance reduction. *Journal of Personality and Social Psychology, 68*, 247-260.

Spellman, B. A., & Holyoak, K. J. (1992). If Saddam is Hitler, then who is George Bush? Analogical mapping between systems of social roles. *Journal of Personality and Social Psychology, 62*, 913-933.

Spellman, B. A., Ullman, J. B., & Holyoak, K. J. (1993). A coherence model of cognitive consistency: Dynamics of attitude change during the Persian Gulf war. *Journal of Social Issues, 49*, 147-165.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261-302). New York: Academic Press.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435-502.

Thibodeau, R., & Aronson, E. (1992). Taking a closer look: Reasserting the role of the self-concept in dissonance theory. *Personality and Social Psychology Bulletin, 18*, 591-602.