Psychological Review
1998, Vol. 105, No. 1, 174–187

0033-295X/98/$3.00

Explorations of Cohen, Dunbar, and McClelland's (1990) Connectionist Model of Stroop Performance

Stephen M. Kanne, David A. Balota, Daniel H. Spieler, and Mark E. Faust
Washington University

The J. D. Cohen, K. Dunbar, and J. L. McClelland (1990) model of Stroop task performance is used to model data from a study by D. H. Spieler, D. A. Balota, and M. E. Faust (1996). The results indicate that the model fails to capture overall differences between word reading and color naming latencies when set size is increased beyond 2 response alternatives. Further empirical evidence is presented that suggests that the influence of increasing response set size in Stroop task performance is to increase the difference between overall color naming and word reading, which is in direct opposition to the decrease produced by the Cohen et al. architecture. Although the Cohen et al. model provides a useful description of meaning-level interference effects, the qualitative differences between word reading and color naming preclude a model that uses identical architectures for each process, such as that of Cohen et al., to fully capture performance in the Stroop task.

J. R. Stroop published his influential articles on attention and interference more than 60 years ago (Stroop, 1935a, 1935b). Since that time, the Stroop task, or variations thereof, has been widely used in many different experimental and clinical settings. The popularity of the Stroop task is in large part due to the robustness of its effects and its potential utility for better understanding such central aspects of human attention as the automaticity of word processing and characteristics of attentional control (see MacLeod, 1991, for a review).

In a standard Stroop task, participants are often exposed to three conditions: congruent, incongruent, and neutral. In the congruent condition, color words (such as *red* and *blue*) are presented in consistent ink colors (e.g., the word *blue* printed in blue). In the incongruent condition, color words are presented in inconsistent ink colors (e.g., the word *blue* printed in red). Both of these conditions are compared with a neutral condition in which the participants name the color of a non-color word or row of Xs. When the task is to name the printed color of the word, and the word is inconsistent with the color it is printed in (the incongruent condition), participants are slower to name the color of that word compared with the neutral condition. This increased color naming latency is called the *Stroop effect*. In addition, when the word is consistent with the color of the word (the congruent condition), there is some facilitation observed, although this effect is smaller than the observed interference effect. Finally, in contrast to the large influence of the word code on color naming performance, there typically is very little influence of the color on word reading performance.

Numerous theories have been proposed to account for the Stroop effect. For example, Stroop originally interpreted his results as suggesting that words evoked only one response, whereas colors evoked more than one response (Stroop, 1935a). Thus, the color code is more susceptible to interference from a distracting stimulus than the word code. More recently, researchers have advanced notions of (a) races between word and color processes (Dyer, 1973; Morton & Chambers, 1973), (b) differences in the degree of automaticity of the two processing pathways (Posner & Snyder, 1975), and (c) independent contributions of color and word processing to performance (Lindsay & Jacoby, 1994).

Cohen, Dunbar, and McClelland (1990) designed a parallel distributed processing (PDP) model of the Stroop task that is very promising in that it incorporates the concepts of relative speed of word and color processing and automaticity within a quantitatively tractable model. According to Cohen et al., the Stroop effect can be captured in a relatively simple connectionist network that assumes that the weights of the connections among color and word pathways vary in a continuous fashion on the basis of differing levels of experience with these two processing dimensions. Because the connection strengths can vary continuously, there is a continuum of automaticity rather than a dichotomy that was viewed as a drawback of earlier accounts of the Stroop task (see MacLeod, 1991, pp. 192–193, for a review). In addition, the Cohen et al. model nicely captures many Stroop phenomena, including the asymmetry between facilitation and interference, stimulus onset asynchrony effects, practice effects, and response set effects.

Our initial aim was to use the Cohen et al. (1990) model as a vehicle to better understand the cognitive changes that Spieler, Balota, and Faust (1996) recently observed in healthy older adults compared with healthy younger adults, and in individuals

Stephen M. Kanne, David A. Balota, Daniel H. Spieler, and Mark E. Faust, Department of Psychology, Washington University. Daniel H. Spieler is now at the Department of Psychology, State University of New York at Binghamton. Mark E. Faust is now at the Department of Psychology, University of South Alabama.

This work was supported by National Institute on Aging (NIA) Grants P01 AGO3991 and RO1 AG10193 and NIA Training Grant AG00030. Portions of this research were presented at the Cognitive Neuroscience meeting, San Francisco, California, in March 1995. Thanks are extended to Jonathan D. Cohen for helpful comments on portions of this research.

Correspondence concerning this article should be addressed to Stephen M. Kanne or David A. Balota, Department of Psychology, Washington University, Campus Box 1125, St. Louis, Missouri 63130. Electronic mail may be sent via Internet to smkanne@artsci.wustl.edu or dbalota@artsci.wustl.edu.

with dementia of the Alzheimer's type, compared with healthy age-matched controls. Spieler et al. argued that the increase in interference that normal older individuals demonstrate on the Stroop task may reflect a breakdown in inhibitory processes. We believed that Spieler et al.'s Stroop results could be interpreted within the Cohen et al. framework to explore the potential loci of the observed inhibitory breakdowns in these populations. In pursuit of this goal, we first attempted to simulate Spieler et al.'s young adult data by using the Cohen et al. model. Recall that the Cohen et al. model used only two color–word responses. Because the experimental data from the Spieler et al. study came from a design that included more than two response alternatives, we extended the architecture of the model to accommodate this difference. However, we found that the extended model encountered some difficulty accounting for the obtained results. This led us to examine the model to determine what produced the difficulty. Our goal is to better understand the implications that these difficulties have for the Cohen et al. model and other similar models that invoke the same type of processing architecture to account for data from two distinct processing dimensions.

This article is divided into five sections. First, we provide a brief overview of the Cohen et al. (1990) model, along with a demonstration that we were able to successfully implement the original model. Second, we demonstrate that the Cohen et al. model encounters some difficulties when applied in a relatively straightforward fashion to the results from Spieler et al. (1996). Third, we explore alternative models with different architectures that may eliminate these problems and argue that a problem in the Cohen et al. model is likely due to the special nature of its architecture in which there are only two responses available. Fourth, we present the results from an experiment that demonstrates problems with the Cohen et al. model when the manipulations of the experiment directly map onto the Cohen et al. architecture. Finally, we conclude with a discussion of the implication of these results for both the Cohen et al. model and other models that attempt to account for two processing dimensions within identical types of processing architectures.

Overview of the Cohen et al. (1990) Model

As shown in Figure 1, the Cohen et al. (1990) model uses a PDP architecture with two main processing pathways: color and word. The model has four input units (two for color and two for word), two output units, and a middle (hidden) layer of four units that lie between the input and the output layers. Moreover, there are two additional input units that encode whether the task is to name the color or to read the word, thus reflecting the role of attention. The two output units correspond to the responses *red* and *green*. When a stimulus is presented to the model, activation spreads through the connections from the input units to the hidden units to the output units. The activation is strictly feed-forward in this model in that there are no reciprocal connections between layers or interconnections within layers.

During training, the weights of the connections between the units, and thus the strength of each pathway, are adjusted by using a back propagation algorithm (see Rumelhart, Hinton, & Williams, 1986). Specifically, during training, an input is presented along with the appropriate output, and through the use of the back propagation algorithm the network gradually decreases the difference between its computed output and the correct output by adjusting the connection weights. To simulate the greater facility in reading words as opposed to naming colors, Cohen et al. (1990) trained the network a greater amount on the word patterns than on the color patterns. The result was that the connections in the word reading pathway became stronger than the connections in the color naming pathway.

After training the network, Cohen et al. (1990) were in a position to determine the adequacy of the model in accounting for actual Stroop performance. On each trial, Cohen et al. first turned on either the word or the color task node. The presentation of this stimulus before any other stimulus is presumably analogous to giving instructions (e.g., "read the word" or "name the color") to participants that allow them to set internal mechanisms for task-appropriate processing. For example, if the task was color naming, the color task node received an activation of one and the word task node received an activation of zero. Next, the corresponding input units for the color and word were activated with ones and zeros. For example, if the stimulus was the word *red* printed in green ink (the incongruent condition), the color unit red would receive a zero, the color unit green would receive a one, the word unit red would receive a one, and the word unit green would receive a zero. The neutral condition (assumed to be a colored row of Xs for color naming) was simulated by activating only the appropriate color or word units after the task units were activated. Once a stimulus configuration was presented, activation propagated through the network during a single cycle. Activation built up gradually across cycles in a cascadic fashion (McClelland, 1979).

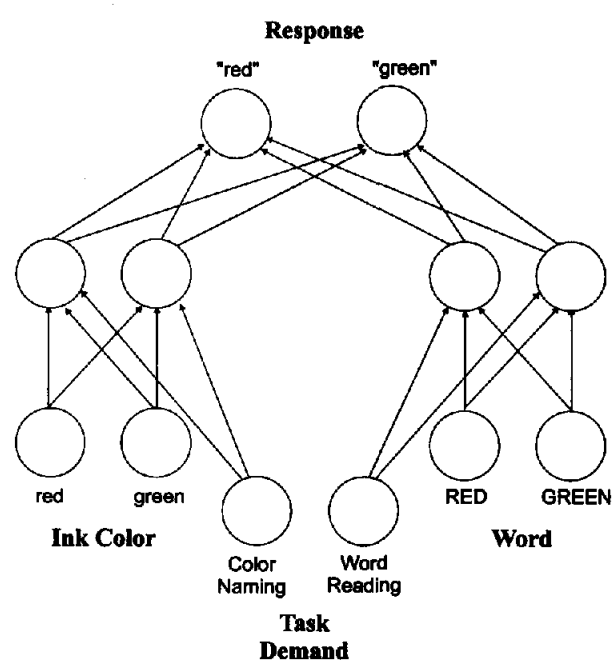

Figure 1. Network designed by Cohen, Dunbar, and McClelland (1990) to simulate Stroop performance (weights are reported in Table 2). From "On the Control of Automatic Processes: A Parallel Distributed Processing Account of the Stroop Effect," by J. D. Cohen, K. Dunbar, and J. L. McClelland, 1990, *Psychological Review, 97*, p. 336. Copyright 1990 by the American Psychological Association. Reprinted with permission of the authors.

Finally, for the network to generate a response, Cohen et al. (1990) associated an evidence accumulator with each of the output units. These evidence accumulators calculated a running total by using previous and current activation values of the output units. When one of the evidence accumulators reached a predetermined threshold, a response was generated. Evidence for a particular response was found by adding the difference between the activation of an output unit and the most active alternative output unit to a running total for that response. Of course, if there are only two output units, as in the Cohen et al. model, the next most active unit is always the alternate response. For example, if the red response unit had an activation of 0.5 and the green response unit had an activation of 0.3, the evidence added to the red response would be 0.2 (0.5 − 0.3, red − green) and the evidence added to the green response would be −0.2 (0.3 − 0.5, green − red). In contrast, when there are more than two response units, the situation is more complex. For example, assume an architecture with three output units: red, green, and blue. If red had an activation of 0.5, green had an activation of 0.3, and blue had an activation of 0.1, then the evidence added to the red response would be 0.2 (0.5 − 0.3, red − green), the evidence added to the green response would be −0.2 (0.3 − 0.5, green − red), and the evidence added to the blue response would be −0.4 (0.1 − 0.5, blue − red). When the evidence for a response reached a predetermined threshold, the network triggered that response as its output.

Two further aspects of the model are noteworthy: First, during each cycle, Gaussian noise was added to the input of each unit and to the response accumulators. Second, all of the hidden units in both pathways were given an initial negative bias. This negative bias has the effect of tonically inhibiting all units in a particular pathway in the absence of input from the attention unit associated with that pathway. The bias was set such that activating the appropriate attention unit gave all of the hidden units in that task pathway an initial net input of zero. Having a net input of zero placed those units in the most sensitive range of their activation function. In contrast, units in the competing pathway were kept at a negative bias and were much less sensitive to their respective inputs. The use of a logistic activation function (see Figure 2) introduces nonlinearity into the model and is credited with imparting many of the model's important behaviors (e.g., Cohen et al., 1990, p. 338). When the net input to a unit is 0.0, large changes in the unit's activation are produced by small changes in the net input. As the net input becomes more positive or more negative, larger changes are needed in the net input to produce similar changes in activation. In mathematical terms, the logistic function's slope is at its peak (around 1) when the net input is 0.0 and decreases as the net input moves away from 0.0 in a positive or negative direction.

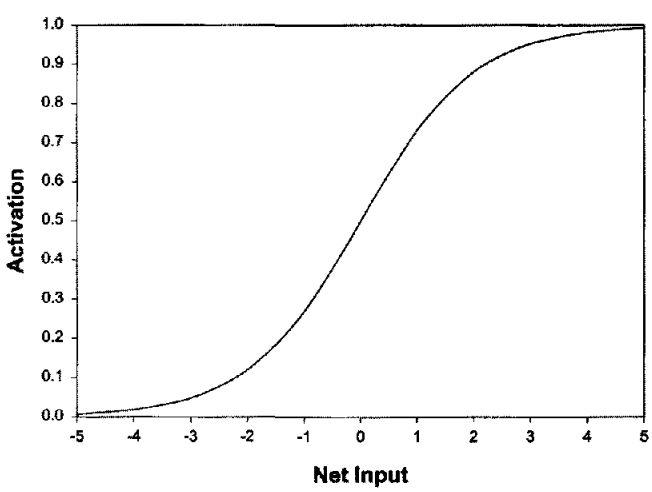

Figure 2. Logistic function used to compute a unit's activation.

Replication of Cohen et al.'s (1990) Simulation

Our first step was to implement the original Cohen et al. (1990) 2-2 architecture displayed in Figure 1. The only difficulty encountered was in the connection weights after training. The weights published by Cohen et al. were ostensibly for a training ratio of 10:1 (Cohen et al., 1990, p. 339, Figure 3). However, the current simulation (Simulation 1) revealed that the weights reported by Cohen et al. after a 10:1 ratio were actually found after training Simulation 1 at a 5:1 ratio. This was later confirmed by Cohen (personal communication, J. D. Cohen, August 18, 1994).

Cohen et al. (1990) did not publish the mean number of cycles to reach a response threshold for each of the conditions. Fortunately, Mewhort, Braun, and Heathcote (1992), who extensively tested Cohen et al.'s model for response time distribution characteristics, published the means from their series of tests. The values that they obtained were found after running 10,000 trials, using Cohen et al.'s published weights. As shown in Table 1, using Cohen et al.'s published weights and 10,000 trials (and using the same values for all of the other parameters), we obtained virtually identical values (considering the variability due to the added Gaussian noise) to those Mewhort et al. reported for color naming. For completeness, Table 1 includes the word reading values, which were not reported by Mewhort et al.

Finally, when the results of Simulation 1 are compared with the data that Cohen et al. (1990) simulated (Dunbar & MacLeod, 1984, Figure 3), an r^2 of .993 is produced, reflecting an excellent fit (Figure 3). The comparison is made by using a simple linear regression analysis, with the simulation's data as the independent variable and the empirical data as the dependent variable, the procedure used by Cohen et al. Thus, we were able to successfully implement the Cohen et al. model.

Modeling the Data From Spieler et al. (1996)

We now turn to our primary goal of attempting to capture the Spieler et al. (1996) results within the Cohen et al. (1990) framework. To match the design in the Spieler et al. study precisely, two changes needed to be made in the Cohen et al. model's architecture. First, instead of two colors and two color words, Spieler et al. used four colors and four color words. Second, because of the qualitatively distinct nature of a row of Xs as a neutral stimulus, Spieler et al. included four non-color words (*bad, poor, deep,* and *legal*) as the neutral stimuli. The use of more stimuli and a different neutral condition required the addition of appropriate input, hidden, and response units. We did not expect this change to alter the ability of the Cohen et al. model to accommodate the results from the Spieler et al.

Table 1
Mean Number of Cycles Obtained by Mewhort et al. (1992) and by Simulation 1 for Color Naming and Word Reading, Using Cohen et al.'s (1990) Published Weights and 10,000 Iterations

	Color naming				Word reading	
	Mewhort et al.		Simulation 1		Simulation 1	
Condition	*M* no. of cycles	*SD*	*M* no. of cycles	*SD*	*M* no. of cycles	*SD*
Congruent	30.3	1.5	30.6	1.5	23.5	0.9
Neutral	35.6	2.2	35.7	2.2	24.0	1.0
Incongruent	48.8	4.9	48.1	4.8	24.7	1.1

study. The implemented architecture (Simulation 2) is displayed in Figure 4.

The amount of training again corresponded to Cohen et al.'s (1990) arbitrary restriction that the model arrive at a correct response in under 50 cycles. The stimuli used to train the model in Simulation 2 (at a 5:1 training ratio) were identical to those used in Simulation 1 but expanded to accommodate the two new colors. In addition, the neutral conditions were simulated by activating a non-color word along with a color input after the appropriate task node was activated, which depended on whether the task was to name the word or the color. The connection weights for Simulation 2 (and all other simulations) are reported in Table 2. The values for the mean number of cycles to reach response criterion based on 1,000 trials are reported in Table 3, and the actual fit of the model to the data is displayed in Figure 5.

As the results show, there is a strong discrepancy between the model's prediction and the observed data. In fact, r^2 is only .55 in Figure 5. The major problem in the model's fit of the data is the difference between word reading and color naming. The model predicts that the word reading and color naming data should be much closer than the empirical evidence indicates. Moreover, as shown by comparing Simulations 1 and 2, one can see that the major difference appears to be that the word reading and color naming latencies are closer in Figure 5. If one attempts to account only for the three color naming data points, the model does an excellent job, producing an r^2 of .999. However, when one includes the word reading data, as Cohen et al. (1990) did, the fit is quite poor.

Search for an Explanation

It is possible that Spieler et al. (1996) used an experimental design that is idiosyncratic and hence not typical of Stroop studies. For example, there were two major characteristics of the Spieler et al. study that were different from the original Cohen et al. (1990) architecture. First, there was an asymmetry between the number of colors (four) and the number of words (eight) in the Spieler et al. study. Second, there were simply more items in the Spieler et al. study (4-8) compared with the Cohen et al. 2-2 architecture. Consequently, one must ask which design is more typical of manipulations within Stroop experiments. In fact, we had difficulty finding Stroop studies with only two colors and two words. More important, we examined the method section of the Dunbar and MacLeod (1984) study, which produced the data that Cohen et al. initially modeled. The Dunbar and MacLeod study included a set of 5 color words that were embedded within 45 non-color words. Thus, the 2-2 architecture implemented by Cohen et al. (see Figure 1) did not accurately reflect the design of the experiment that it simulated. If an attempt is made to imple-

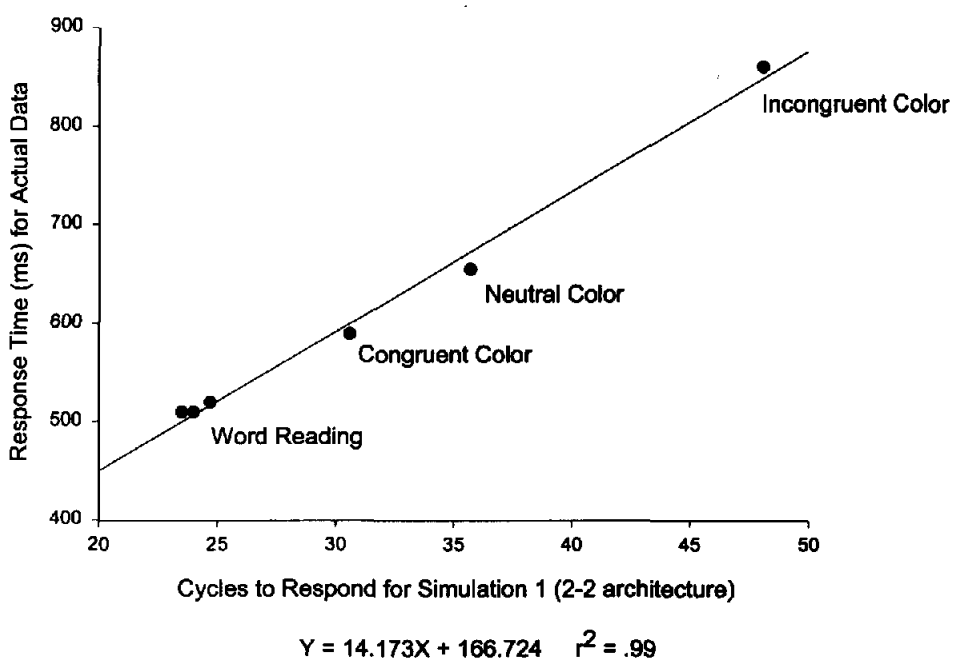

Figure 3. Regression plot for Simulation 1 and data from Dunbar and MacLeod (1984, Figure 3, p. 630).

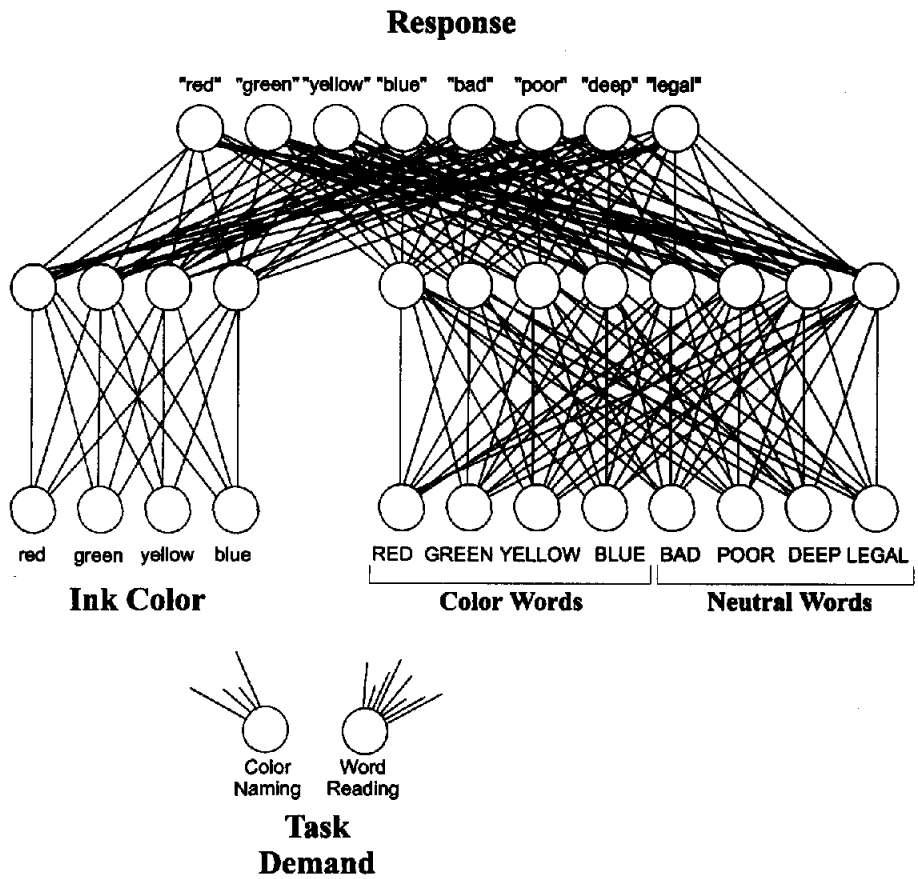

Figure 4. The architecture used in Simulation 2 to simulate Spieler et al.'s (1996) data (weights are reported in Table 2).

Table 2
Connection Weights (and Number of Connections) and Resting Activations of Output Units for Each Simulation

Pathway	Simulation 1: 2-2	Simulation 2: 4-8	Simulation 3: 3-3	Simulation 4: 4-4
Color				
Input to hidden-excitatory	2.2 (1)	2.6 (1)	2.3 (1)	2.5 (1)
Input to hidden-inhibitory	−2.2 (1)	−1.8 (3)	−2.2 (2)	−2.1 (3)
Hidden to output-excitatory	1.3 (1)	1.7 (1)	1.5 (1)	1.8 (1)
Hidden to output-inhibitory	−1.3 (1)	−1.3 (3)	−1.4 (2)	−1.5 (3)
Response unit resting activation	.50	.20	.32	.18
Word				
Input to hidden-excitatory	2.6 (1)	3.3 (1)	2.9 (1)	3.1 (1)
Input to hidden-inhibitory	−2.6 (1)	−1.9 (7)	−2.5 (2)	−2.3 (3)
Hidden to output-excitatory	2.5 (1)	3.9 (1)	2.9 (1)	3.3 (1)
Hidden to output-inhibitory	−2.5 (1)	−1.9 (7)	−2.6 (1)	−2.6 (3)
Response unit resting activation	.50	.01	.23	.09

Note. In each architecture, each input unit has one excitatory connection to the hidden layer, and the remainder are inhibitory. The hidden layer unit also has one excitatory connection to the response layer, and the rest are inhibitory. For example, in the 4-4 architecture, the *red* word input unit excites one hidden layer unit (3.1) and inhibits the three other hidden units (−2.3). The hidden unit receiving the excitatory input excites the "red" output unit (3.3) and inhibits the other three responses (−2.6).

Table 3
Mean Number of Cycles for Color Naming and Word Reading Obtained by Simulation 2

	Color naming		Word reading	
Condition	*M* no. of cycles	*SD*	*M* no. of cycles	*SD*
Congruent	25.4	0.9	28.9	0.9
Neutral	34.5	2.5	31.0	2.5
Incongruent	49.0	7.6	31.3	7.6

ment architectures that more faithfully reflect the designs used to examine Stroop performance, then it appears that some difficulties will be encountered within the Cohen et al. modeling framework.

Asymmetry

One of the two major differences between the Cohen et al. (1990) model and Simulation 2 has to do with the symmetry of their architectures. Initially, the change in symmetry appeared to be a promising candidate for the observed differences. That is, only the use of a symmetrical design with two input and two hidden units for each pathway (such as the Cohen et al. model) will result in a resting activation of 0.5 for the response units when either task unit is activated. In this case, input to the response units sums to 0.0, reflecting a balance of excitatory and inhibitory input. In an asymmetrical design, this balance is not achieved. When the task unit for the color pathway is activated in Simulation 2, all of the response units that have corresponding units in the color pathway (red, green, yellow, and blue) have an activation level of 0.2. When the task unit for the word pathway is activated, all of these response units have an activation of 0.01. The disparity between these resting activations is produced by the larger amount of inhibitory input to the response units from the word pathway versus the color pathway. If the Cohen et al. model was extended to include more units, but remained symmetrical, one could examine whether the asymmetry was the cause of the problems.

Table 4
Mean Number of Cycles for Color Naming and Word Reading Obtained by Simulation 3 and Simulation 4

	Simulation 3		Simulation 4	
Condition	*M* no. of cycles	*SD*	*M* no. of cycles	*SD*
Color naming				
Congruent	29.0	1.2	28.1	1.1
Neutral	34.3	1.8	33.5	1.7
Incongruent	47.9	5.1	48.2	5.2
Word reading				
Congruent	23.5	0.9	24.4	0.8
Neutral	24.1	0.9	25.3	0.9
Incongruent	24.5	0.9	25.8	1.0

To address this issue, we implemented two new architectures that had more units yet preserved symmetry. We used similar architectures to that in Simulation 2, but without the non-color words. Therefore, these simulations had the same architecture as the Cohen et al. (1990) model, but extended to three (Simulation 3) and four (Simulation 4) colors. The values of mean number of cycles for each condition that were obtained (for 1,000 trials) are shown in Table 4. Because Simulation 4 used four colors, as did Spieler et al. (1996), these values were analyzed against the Spieler et al. data. The fit is somewhat better than in the preceding simulation (r^2 = .859, see Figure 6), because the word condition is now a bit faster than in the

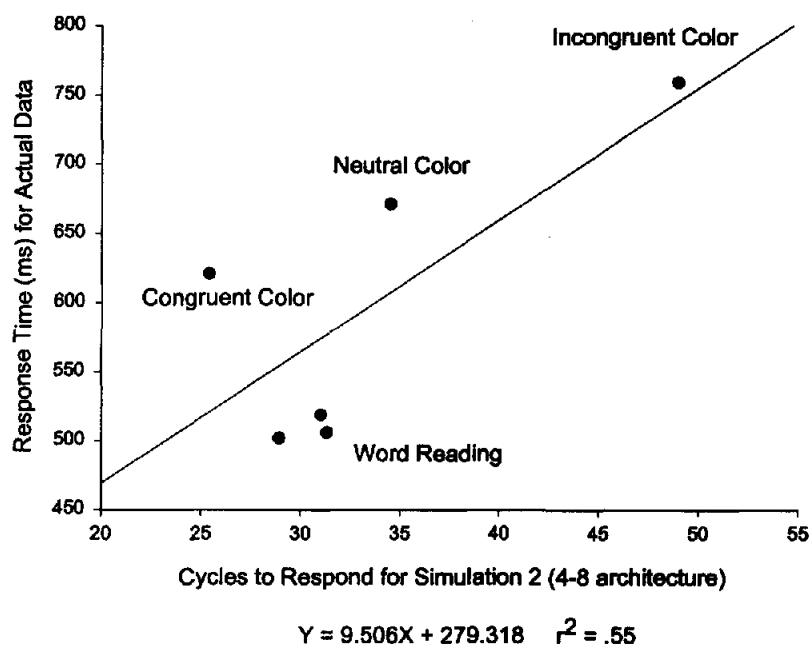

Figure 5. Regression plot for Simulation 2 and the Spieler et al. (1996) data for young participants.

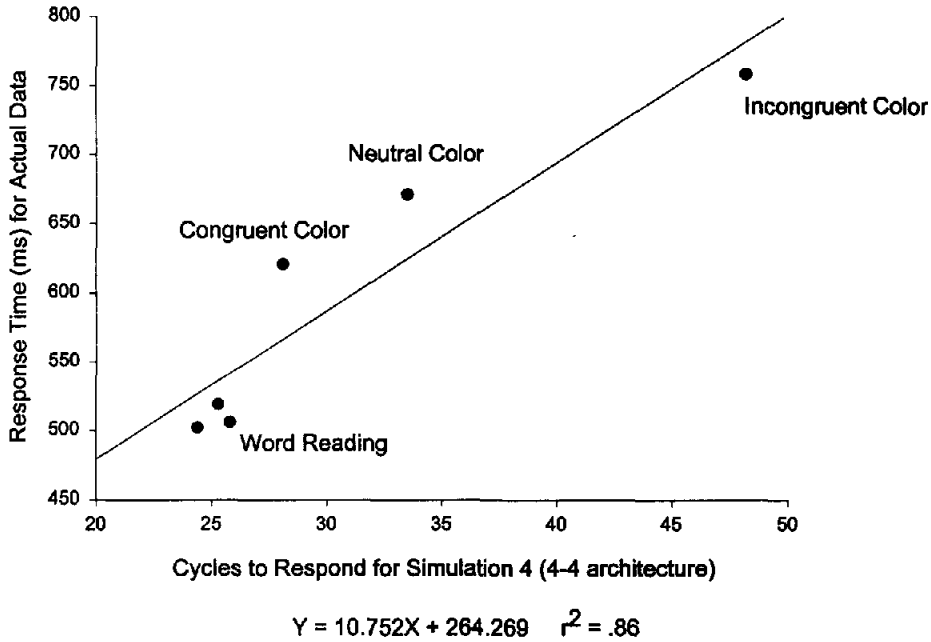

Figure 6. Regression plot for Simulation 4 and the Spieler et al. (1996) data for young participants.

previous simulation. However, as shown in Figure 6, the model's fit remains poor even with a symmetric architecture. Thus, symmetry does not eliminate the problem in accounting for the empirical data.[1]

Set Size

The second major difference between the Cohen et al. (1990) 2-2 architecture and the Spieler et al. (1996) study, along with most Stroop studies, is a difference in set size. Thus, we explored the possibility that differences in set size may be contributing to the observed discrepancy between the model and the empirical data. Recall that the main difficulty encountered when trying to account for the empirical data is that the model exhibits a decrease in the relative difference between the word and color conditions as the number of units increases across architectures. Figure 7 shows that the decrease in the relative difference is the result of the word response latencies increasing as the number of units in the architecture increases, whereas the color response latencies actually decrease as the number of units in the architecture increases. This pattern appears inconsistent with that found in the empirical data (see MacLeod, 1991, p. 177, for a review). Specifically, the time to name colors typically increases, whereas the time to read words is only slightly affected. Figure 8 displays the results from an experiment (reported in detail in the section Experimental Implementation of the Simulations) where the number of colors increased from two to four, just as in the simulations.

To better understand why the number of cycles needed to respond in the word and color conditions converges as one increases the number of response alternatives, it is important to note how the weights are distributed in the various network architectures. In the Cohen et al. (1990) model, the connection weights to each response unit or hidden unit are of equal and opposite sign. For example, in the Cohen et al. model, the weight in the word pathway from one hidden unit to the red response unit was 2.5 and to the green response unit was −2.5. The weights in the color pathway were 1.3 and −1.3. Thus, the connection weights were balanced, and the response unit was placed in the most sensitive range of activation at 0.5 (see Figure 2). However, in architectures that have more than two units in each pathway, the excitatory and inhibitory inputs to the response units are not balanced. For example, in the 4-4 architecture, the weight in the word pathway to the red response unit from one hidden unit was 3.3, whereas the weights were −2.6 from each of the other three hidden units. In the color pathway, there was one weight of 1.8 and three weights of −1.5. The response units are receiving a larger amount of inhibitory input compared with excitatory input in architectures with more than two units in each pathway. Thus, there is not the connection weight balance, so the response units are not placed in the most sensitive 0.5 range. The pattern of weight distribution produced by an increase in the number of units in each pathway appears to contribute to the convergence of the word and color conditions.

The convergence of the color and word conditions in the symmetric architectures appears to be caused by the same mechanism responsible for the difficulties in the asymmetric architecture. One way of examining this effect is by looking at the

[1] It should be noted that Simulation 4's architectural design does not exactly match Spieler et al.'s (1996) empirical design that it simulated, which makes interpreting the results more difficult. Simulation 4's design would be analogous to using a row of Xs in the neutral condition instead of the non-color words used in the Spieler et al. experiment. However, this does not account for the problems with the model, because we later present empirical data from a study that includes a neutral row of Xs and demonstrate that the problems still remain.

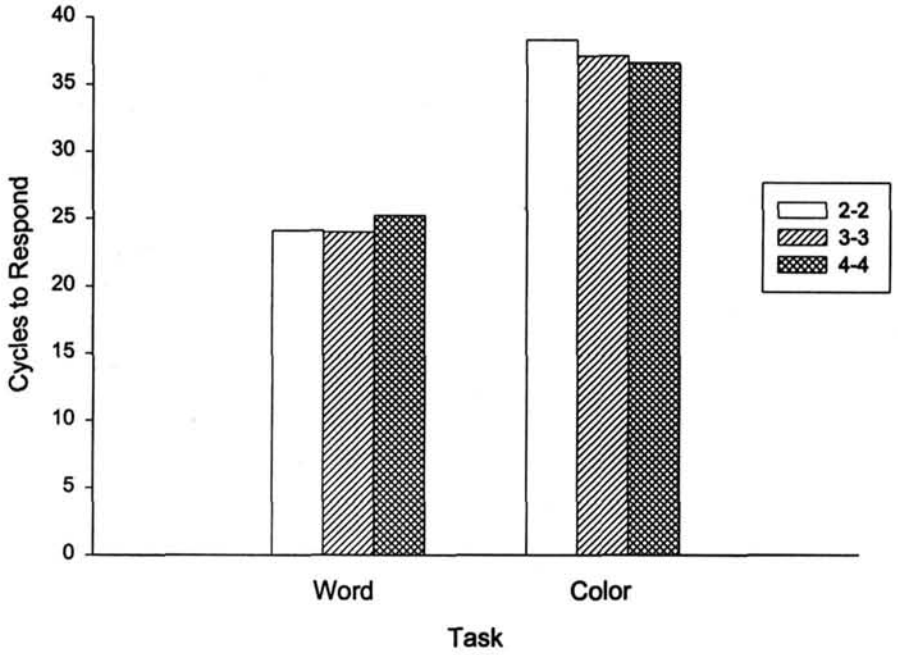

Figure 7. Number of cycles to respond for word and color (congruent, incongruent, and neutral conditions averaged) for the 2-2, 3-3, and 4-4 simulation architectures.

"resting" activation of the response units after the attention units are activated. In the Cohen et al. (1990) 2-2 model, after either color or word task unit was activated, the response units attained a resting activation of 0.5. This value was arrived at by the summed activation of 0.0 coming from the balanced inputs. This fact is important for two reasons: First, the response units are in the most sensitive range of their activation function, allowing them to respond quickly to small changes in activation (see Figure 2). Second, both response units begin at 0.5, which allows their activations to achieve the greatest separation in the shortest amount of time by moving in opposite directions. The speed of separation between the two units represents response

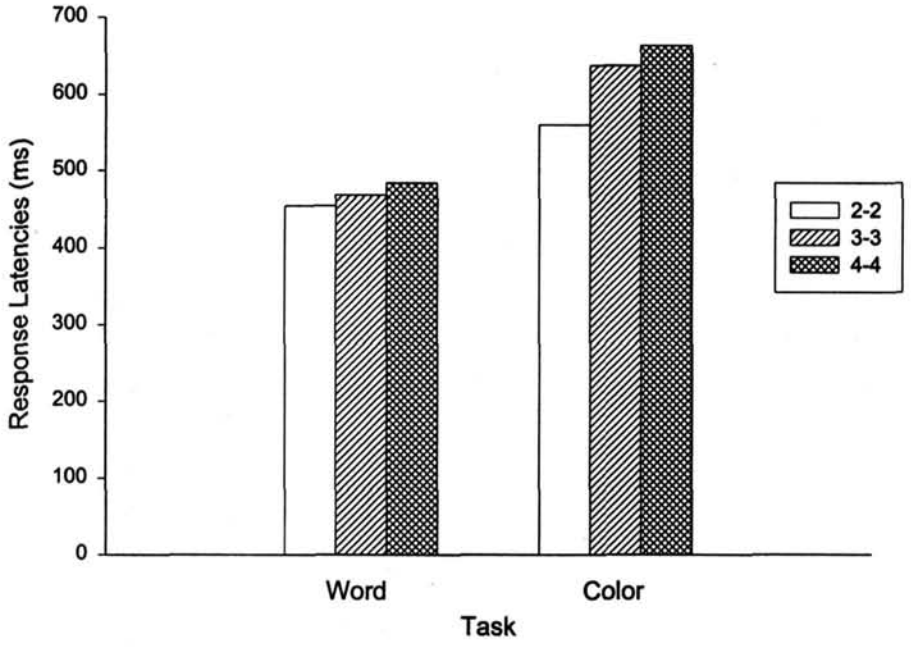

Figure 8. Response latencies for word and color (congruent, incongruent, and neutral conditions averaged) for the 2-2, 3-3, and 4-4 set sizes from the experiment.

unit competition and determines the speed of the response. Therefore, if both units begin at some value closer to the activation extremes (0.0 or 1.0), the response will be slowed. For example, consider the different responses across networks when the word task is activated. In the 4-4 architecture, each response unit attained a resting activation of 0.09 after activation of the word task unit. The lower value was a direct result of the greater amount of inhibitory input (one weight of 3.3 and three weights of −2.6). In the asymmetrical 4-8 architecture (four colors and eight words), each color response unit attained a resting activation of 0.01 after the word task unit was activated (one weight of 3.9 and seven weights of −1.9). The decreased activations move these response units into a less sensitive range and reduce the ability of the units' activations to separate quickly, which results in slower responses. When the color task unit is activated, a decrease in resting activation also takes place, but to a reduced extent because of the weaker weights in the color pathway (see Table 2). Therefore, as the number of units increases, the number of cycles to respond will increase in the word condition relative to the color condition, creating a convergence of word and color performance.

In the present simulations, although the word reading latencies increased, the color naming latencies did not increase as the argument above would indicate. Recall that the networks received increased training, which was necessary to allow all conditions to respond correctly in under 50 cycles. Therefore, as network size increased, the incongruent color condition was held relatively constant at around 50 cycles. The neutral color condition became slightly faster because of the stronger weights in the color pathway that result from increased training. The stronger weights in the color pathway, as well as the stronger weights in the word pathway, contributed a greater amount of activation and thus sped the response in the congruent color condition.

Perhaps a more appropriate method of comparing results across architectures may be to lift the arbitrary 50-cycle restriction and instead train all the networks to an equal degree.[2] This assumption was tested by training Simulation 3 (3-3 architecture), Simulation 4 (4-4 architecture), and Simulation 2 (4-8 architecture) at 1,000 epochs, the same as Simulation 1 (2-2 architecture). As can be seen from Table 5, the problem of the word and color conditions converging is not eliminated. The number of cycles needed to respond in the word conditions increased across the simulations as the number of units increased, whereas the color-congruent condition remained relatively stable. Again, the word conditions approached, and eventually overtook, the color-congruent condition in a manner that is incompatible with empirical findings.[3]

In summary, the greatest separation between the word and color conditions is achieved with a symmetric architecture with the fewest possible units: the 2-2 architecture used by Cohen et al. (1990). This greater separation produces the best fit to the empirical Stroop data that Cohen et al. modeled. However, if one uses an architecture that directly reflects the empirical manipulations that are being modeled by adding additional units, the model is less able to account for the empirical data because of the convergence of the word and color response latencies.

Table 5
Mean Number of Cycles for Color Naming and Word Reading for the 2-2 (Simulation 1), 3-3 (Simulation 3), 4-4 (Simulation 4), and 4-8 (Simulation 2) Architectures When Each Was Trained at 1,000 Epochs

Condition	2-2	3-3	4-4	4-8
Color naming				
Congruent	30.6	30.5	31.1	30.7
Neutral	35.7	36.8	38.3	50.9
Incongruent	48.1	54.1	59.2	107.6
Word reading				
Congruent	23.5	24.0	25.6	30.7
Neutral	24.0	24.6	26.3	32.2
Incongruent	24.7	25.1	26.8	32.3

Experimental Implementation of the Simulations

As noted earlier, it appears that increasing set size (the number of color and word units) decreases the difference between

[2] Another possibility is to increase the training ratio, which would speed the word responses and possibly alleviate the color–word convergence. This possibility was tested by training the 4-4 architecture, which was trained at a 5:1 ratio, at a 10:1, and at a 20:1 ratio. To allow for direct comparisons, the number of epochs trained was adjusted for each ratio such that the color pathway always received 300 epochs. The manipulation succeeded in speeding the word condition. For example, in the neutral word condition, it took 25.3 cycles to respond for the 5:1 ratio, 24.0 cycles for the 10:1 ratio, and 23.2 cycles for the 20:1 ratio. However, as the ratio increased, the interference effects became highly exaggerated: 48.2 cycles to respond for the 5:1 ratio, 75.0 cycles for the 10:1 ratio, and 466.4 cycles for the 20:1 ratio. The neutral color condition remained constant for each ratio (33.5 cycles), whereas the congruent color condition became faster: 28.1 cycles to respond for the 5:1 ratio, 27.2 cycles for the 10:1 ratio, and 26.2 cycles for the 20:1 ratio. This pattern of results led to a decrease in the model's ability to match the empirical results as the training ratio was increased: $r^2 = .88$ for the 5:1 ratio, $r^2 = .80$ for the 10:1 ratio, and $r^2 = .69$ for the 20:1 ratio.

[3] A suggestion by Jonathan D. Cohen (personal communication, November 7, 1996) prompted another way of examining set size manipulations. The notion is to train a single network with a large number of units and then modulate attention preferentially to a small subset of the units in the hidden layer. The size of the subset will simulate the various set sizes. A larger network would be more similar to an actual human participant's color–word knowledge structure. These simulations were implemented by using the trained 4-4 and 4-8 networks and allowing the attention task units to influence only two of the hidden units in the relevant pathway. The results were then compared with the empirical findings from the reported 2-2 experiment. The results were disappointing ($r^2 = .90$ for the 4-4 network and $r^2 = .77$ for the 4-8 network), with the same pattern of color–word convergence occurring. Again, the resting activations were placed in a less sensitive range and moved closer to an activation extreme. Interestingly, in these simulations, the resting activations increased rather than decreased as before. However, the same mechanism responsible for the convergence in the previous simulations operated in these simulations. For example, in the 4-4 network, when two units were used and the word task unit was activated, the response units attained a resting activation of .59. When the color task unit was activated, the response units attained a resting activation of .54. Thus, the word pathway began in a less sensitive range and closer to an activation extreme. The imbalance in the 4-8 network was worse with only two units used in the response set. The resting activation when the word task unit was activated was .73, and it was only .55 when the color task unit was activated, which decreased the fit.

word and color performance in the Cohen et al. (1990) model. There is some evidence to suggest that increasing set size has the opposite effect in experimental studies of Stroop performance (see MacLeod, 1991, p. 177). Specifically, increasing set size appears to slow down color naming and has little influence on word reading, thereby increasing the difference between color and word performance. Unfortunately, however, we are unaware of any study that systematically manipulated set size, without the introduction of additional variables. In fact, we were unable to find any Stroop study that directly matched the 2-2 architecture that Cohen et al. implemented. Therefore, we decided to conduct an experiment that factorially crossed set size to determine whether the results of such an experiment could be captured by the Cohen et al. model.

Our experiment included three separate set sizes. One set size had two response alternatives and matched the Cohen et al. (1990) model's architecture. The remaining two set sizes had three and four response alternatives to match the 3-3 and 4-4 architectures, respectively (see the Appendix for a full discussion of the experimental method and results). The results from this experiment can be used in two ways: First, the data derived from the 2-2 set size can be used to directly test the Cohen et al. architecture. Second, the data derived across the set size manipulation can be used to directly test the influence of set size on the word–color difference.

The mean response latencies for both word reading and color naming for each set size are presented in Table 6. Note that word reading is faster than color naming for each set size. More important, however, as one compares the difference between word reading and color naming across set sizes, it is quite clear that increasing the number of response alternatives slows down word reading latencies slightly, whereas it slows down color naming latencies to a much greater extent. Specifically, the 2-2 set size resulted in a word–color difference of 105 ms, the 3-3 set size resulted in a word–color difference of 169 ms, and the 4-4 set size resulted in a word–color difference of 179 ms. This is precisely the pattern suggested by MacLeod (1991) and is opposite the pattern observed when one increases set sizes in the Cohen et al. (1990) model.

In addition to the influence of set size on word–color differences, there is also an intriguing pattern in the facilitation and interference effects. Specifically, there is relatively little influence of word–color congruency for word reading performance, precisely as the Cohen et al. (1990) architecture predicts. However, when one considers the color naming performance, there appears to be an interesting trade-off between facilitation and interference effects. Specifically, there are large interference effects that increase across set sizes and comparatively small facilitation effects that are eliminated, and actually reversed, at set size 3-3 and set size 4-4. In fact, overall, there is not a reliable facilitation effect in color naming performance. This is not terribly surprising because, as MacLeod (1991) suggested, there is relatively little facilitation when one uses a neutral row of Xs as a baseline to measure facilitation and interference effects. Unfortunately, however, this is troublesome for the Cohen et al. architecture because each of the implemented architectures produced facilitation effects, albeit smaller than the interference effects. Of course, it is possible that we simply used an inappropriate neutral baseline condition. However, consistent with that suggested by the Cohen et al. architecture, we used the row of Xs as a neutral baseline. If one used another baseline condition, such as the neutral non-color words used in the Spieler et al. (1996) study, then one would encounter the other difficulty of an asymmetric architecture and increasing set size that initially led to a number of the present concerns. The point is that if one conducts an experiment that mimics the implemented architecture, then one finds little facilitation in the data but clear facilitation in the model. If one turns to a different neutral baseline, and actually implements this alternative neutral baseline, then one encounters problems of asymmetry and the problems associated with increasing set size.

Table 6
Mean Response Latencies for Color Naming and Word Reading for Each Set Size in the Experiment

	2-2 set size		3-3 set size		4-4 set size	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*
Color naming						
Congruent	529	63	596	74	618	83
Neutral	534	71	586	72	610	72
Incongruent	615	87	728	131	761	97
Word reading						
Congruent	449	52	466	53	481	61
Neutral	452	52	466	54	484	66
Incongruent	461	61	471	56	488	62

The primary motive for conducting this experiment was to see how well the Cohen et al. (1990) architecture fits the empirical data that directly matched the experimental design. As shown in Figure 9, the overall fit is relatively poor (r^2 = .716) when all of the empirical data from each set size and each model are used in the analysis.[4] One might argue that because different models and different participants were used in each design, this may underestimate the quality of fit for a given implemented architecture. However, individual analysis of each simulation versus its analogous empirical data reveals that as the number of units increases, despite being symmetric, the quality of the fit systematically decreases (r^2 = .96 for the 2-2 architecture, r^2 = .92 for the 3-3 architecture, r^2 = .88 for the 4-4 architecture, and r^2 = .55 for the 4-8 architecture).

General Discussion

The original goal of this study was to model the changes found in normal aging and dementia of the Alzheimer's type on Stroop task performance in the Spieler et al. (1996) study by using Cohen et al.'s (1990) connectionist model. To accomplish this goal, we implemented a slightly different architecture that corresponded to the conditions used in the Spieler et al. study. Although this new architecture captured the color naming data, it was unable to account for the overall data pattern when word reading data were included. To address the question of whether

[4] We also submitted the results from the simulations when each was trained at 1,000 epochs (see Table 5) to a regression analysis. The result was an r^2 of .67, confirming the conclusion that the color–word convergence was occurring and producing a poor fit to the empirical data.

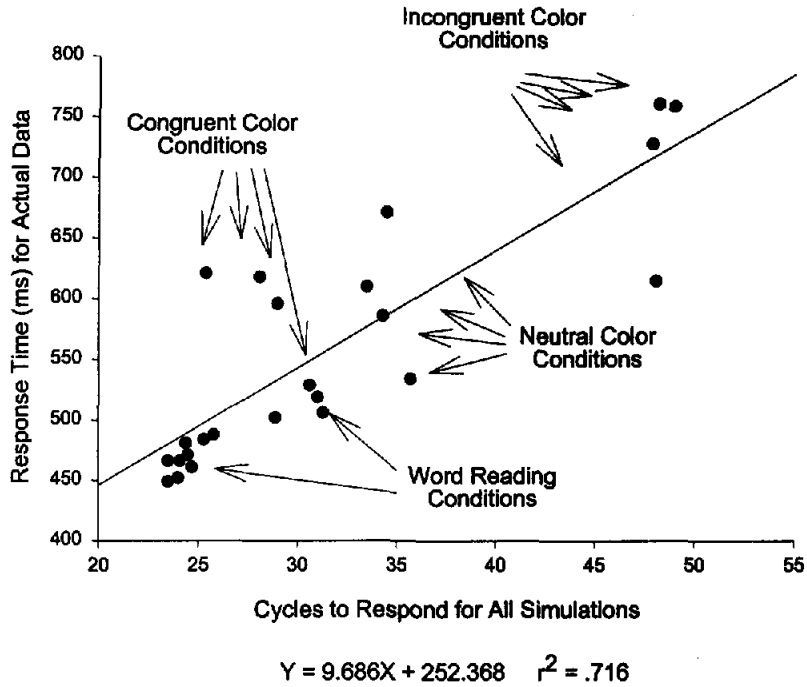

Figure 9. Regression plot for Simulations 1, 3, and 4 (color naming and word naming) and the results from the 2-2, 3-3, and 4-4 set size manipulations from the experiment.

the problems found were the result of using an asymmetrical architecture, we implemented two extended but symmetric architectures. These symmetric architectures produced better fits to the empirical data but encountered problems in capturing the individual patterns of results from the various conditions. These failures caused us to reexamine the data the Cohen et al. model simulated and raised the question of whether the implemented architecture mimicked the manipulations in the targeted experiment. We observed that the experiment simulated by Cohen et al. (i.e., Dunbar & MacLeod, 1984) involved a design that included more items in the response set than the 2-2 architecture implemented in the Cohen et al. model. To test the implications of this discrepancy, we reported the results from an experiment that directly investigated set size differences. The Cohen et al. architecture was unable to capture important aspects of the results from this experiment.

Cohen et al. (1990) emphasized that the

> model shows that two processes that use qualitatively identical mechanisms and differ only in their strength can exhibit differences in speed of processing and a pattern of interference effects that make the process look as though one is automatic and the other is controlled. (Cohen et al., 1990, p. 334)

Although there is little doubt that the model does in fact show these effects, the results are based on the stated assumption that the word reading and color naming processes use "qualitatively identical mechanisms." We believe that the present analyses and the extant literature question this assumption. Thus, the model will encounter difficulties when attempting to account for both word reading and color naming aspects of Stroop performance.

It is likely that there are many different processes involved in word reading and color naming. For example, we are sympathetic to arguments that have been long espoused by word recognition researchers suggesting that visual word reading can be accomplished through a pathway that involves computations that map phonological information directly from the orthographic pattern (e.g., Coltheart, 1978; Seidenberg & McClelland, 1989). In fact, such a nonsemantic route in word reading may be even more prevalent when the stimuli are repeated within the experimental session as in most Stroop studies. This additional route would keep the word reading response latencies relatively fast and impervious to set size manipulations and inconsistent color information, as indicated in the observed data. Although it may be possible to use a nonsemantic route with sufficient practice, we believe that such a route is much less likely to play a role in color naming. In color naming, participants are more likely to rely on the hue of the stimulus to access some meaning-level representation, which in turn drives the relevant name code. The additional analyses involved in color naming will keep performance in this task relatively slow compared with word reading. Clearly, this will in part depend on set size. When one is required to use more color responses within an experiment, this will increase the likelihood that participants cannot rely on simple stimulus–response mapping to generate the color name. Thus, one would expect a considerable slowdown in color naming as one increases the set size. Because words are more likely to directly drive the response through consistent orthographic to phonologic mapping, an increase in set size will produce a smaller influence on performance, as the present results indicate.

The point of this discussion is not to present a model of Stroop interference. There are many such models available that have been proffered with varying degrees of success (see Mac-

Leod, 1991, pp. 187–193, for a review). We believe that there are sufficient differences between the processing pathways of word reading and color naming that a model that uses identical architectures for each will be unlikely to fully capture the operations involved in both tasks within Stroop studies. In this light, it appears that the Cohen et al. (1990) model may be a useful model of the meaning-level interference effects that occur in Stroop studies, above and beyond the task-specific pathways that are relevant to word reading and color naming. However, without the implementation of these additional task-specific operations, we believe that it is unlikely that one will be able to fully capture the characteristics of word reading and color naming in the Stroop task. There are simply too many differences between the operations involved in word reading and color naming to fully capture differences in performance by using a model with identical architectures for each task.

The observed problems may simply be a result of the particular implementation that Cohen et al. (1990) used in accounting for Stroop performance or these problems may be more fundamental with respect to the underlying theoretical assumptions. Cohen et al's theoretical approach relies on the assumption that differences in the amount of practice between word reading and color naming are the primary determinants of the pattern of data obtained in studies of the Stroop effect. We have addressed implementational issues related to set size, asymmetry, and training to better understand the observed problems. However, we have been unable to overcome the discrepancy between the prediction of the model and the observed data regarding differences in word reading and color naming performance. Hence, we believe that the observed difficulties arise from a problem with Cohen et al.'s theoretical assumption that word reading and color naming use qualitatively identical processing pathways and that the primary difference between the pathways is the degree of practice. The breakdown in the model's fit to the empirical data was primarily the result of the inability of the simulations to accurately reflect speed differences in word conditions relative to color conditions. It is important to note that the model did do a much better job when only color naming performance was considered. However, in the Cohen et al. model, the two tasks affect each other in proportion to the level of automaticity each task has attained. Therefore, it is impossible to consider only one task without having implications for the other.

As noted above, we believe that it will be necessary to implement a qualitatively distinct route for word reading. A separate direct route for mapping the visual stimulus word onto an articulatory output may be the most obvious way to account simultaneously for both the asymmetric congruency effects in word reading and color naming and the word advantage over color naming (see Phaf, Van Der Heuden, & Hudson, 1990). However, an area of investigation that deserves further exploration is the use of interactive models that maintain identical architectures for the word reading and color naming pathways. These models use excitatory between-layer connections and inhibitory within-layer connections (see Cohen & Huston, 1994). However, for these interactive models to be successful, they will have to demonstrate that they can overcome the difficulties of convergence, in addition to successfully capturing all of the effects that the original noninteractive Cohen et al. (1990) model captured, such as learning and attentional phenomena.

Despite these weaknesses, we should emphasize that we believe the Cohen et al. (1990) model has many strengths. For example, as noted, we believe that it serves as an excellent framework to account for the meaning-level interference effects that occur in Stroop performance. It is also one of the few computational models of Stroop performance available and, therefore, provides a useful framework for empirical tests. In addition, the model captures many aspects of color naming in the Stroop task, such as the asymmetry between facilitation and interference, stimulus onset asynchrony effects, practice effects, and response set effects. Finally, the inclusion of a task module that maintains instructional set is an important characteristic of the model that is often neglected in cognitive models. The major criticism that we have uncovered is that using identical architectures for both word reading and color naming routes may not be sufficient to accommodate latencies for each task. We do not believe that the processes in these two routes are "qualitatively identical," and we believe that eventually the unique characteristics of each route will need to be implemented for a model to fully capture the wealth of 60 years of empirical work on the Stroop effect.

References

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

Cohen, J. D., & Huston, T. A. (1994). Progress in the use of parallel distributed processing models for understanding attention and performance. In C. Umilta & M. Moscovitch (Eds.), *Attention and performance: XV. Conscious and nonconscious information processing* (pp. 453–476). Cambridge, MA: MIT Press.

Coltheart, M. (1978). Lexical access in a simple reading task. In G. Underwood (Ed.), *Strategies of information processing* (pp. 151–216). London: Academic Press.

Dunbar, K., & MacLeod, C. M. (1984). A horse race of a different color: Stroop interference patterns with transformed words. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 622–639.

Dyer, F. N. (1973). The Stroop phenomenon and its use in the study of perceptual, cognitive, and response processes. *Memory & Cognition, 1*, 106–120.

Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process dissociations: The relationship between facilitation and interference. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 219–234.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrated review. *Psychological Bulletin, 109*, 163–203.

McClelland, J. L. (1979). On the time-relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86*, 287–330.

Mewhort, D. J., Braun, J. G, & Heathcote, A. (1992). Response time distributions and the Stroop task: A test of the Cohen, Dunbar, and McClelland (1990) model. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 872–882.

Microsoft QuickBASIC 4.5 [computer software]. (1988). Seattle, WA: Microsoft Corporation.

Morton, J., & Chambers, S. M. (1973). Selective attention to words and colours. *Quarterly Journal of Experimental Psychology, 25*, 387–397.

Phaf, R. H., Van Der Heuden, A. H. C., & Hudson, P. T. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology, 22*, 273–341.

Posner, M. I., & Snyder, C. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition* (pp. 55–85). Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructures of cognition* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Seidenberg, M., & McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523–568.

Spieler, D. H., Balota, D. A., & Faust, M. E. (1996). Stroop performance in normal older adults and individuals with senile dementia of the Alzheimer's type. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 461–479.

Stroop, J. R. (1935a). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–661.

Stroop, J. R. (1935b). The basis of Ligon's theory. *American Journal of Psychology, 47*, 499–504.

Appendix

Experimental Method and Results

Method

Participants

A total of 72 participants took part in the experiment. They were paid $5 each. All of the participants were recruited from the undergraduate student population at Washington University.

Design

The experiment was a 2 (order: word reading block first vs. color reading block first) × 3 (set size: 2-2 vs. 3-3 vs. 4-4) × 2 (task: word reading vs. color naming) × 3 (condition: congruent vs. neutral vs. incongruent) mixed-factor design. Order and set size were the only between-participant factors. The primary dependent measure was response latencies to the target word or color.

Apparatus

An IBM compatible computer was used to display the stimuli on an NEC Multisynch 2A 14 inch video graphics adaptor color monitor and to collect response times. Voice onset latency was measured by interfacing a Gerbrands Model G1341T voice-operated relay with the computer.

Materials

The word stimuli consisted of four color names (*red, blue, green,* and *yellow*) and a row of Xs. The color stimuli consisted of the blue, green, red, and yellow colors available from Microsoft QuickBASIC 4.5 (Microsoft QuickBASIC 4.5 colors 1, 2, 4, and 14; Microsoft QuickBASIC 4.5, 1988). The stimuli were presented in 40-column mode and were subtended 1.5° to 3° of visual angle.

Each participant received either the 2-2 set size, the 3-3 set size, or the 4-4 set size. The experiment consisted of two tasks, word reading and color naming, each of which included congruent, incongruent, and neutral trials. The color naming task included neutral stimuli, which consisted of a row of Xs displayed in the various colors. The word reading task included neutral stimuli, which consisted of a color word printed in white. As noted, task order was counterbalanced across participants within set size.

To permit direct comparisons across the set sizes, it was necessary to equate repetition effects. For example, if the total number of trials was the same for the 3-3 set size and the 2-2 set size, then participants in the 2-2 set size would have received more repetition on each individual stimulus. Therefore, the amount of exposure to each stimulus was held constant over the experiments, which changed the total number of trials across experiments.

Participants performed the color naming and word reading tasks in separate blocks. Within each block, the stimuli were chosen from the four possible words and the four corresponding colors. The 2-2 set size used only two of the four possible words for a given participant, the 3-3 set size used three words, and the 4-4 set size used all four words. In the congruent condition, each color word appeared 12 times in its corresponding color (2 colors × 12 presentations for the 2-2 set size, 3 colors × 12 presentations for the 3-3 set size, and 4 colors × 12 presentations for the 4-4 set size) In the incongruent condition, each color word appeared 12 times in nonmatching colors (2 colors × 12 presentations for the 2-2 set size, 3 colors × 12 presentations for the 3-3 set size, and 4 colors × 12 presentations for the 4-4 set size). In the neutral condition, each color word appeared 12 times in white or a row of Xs appeared in each of the two colors (2 color words or colored Xs × 12 presentations for the 2-2 set size, 3 color words or colored Xs × 12 presentations for the 3-3 set size, and 4 color words or colored Xs × 12 presentations for the 4-4 set size). This produced a total of 72 trials for each block in the 2-2 set size, 108 trials for each block in the 3-3 set size, and 144 trials for each block in the 4-4 set size.

The same colors were used for both blocks for each participant. When fewer than four colors were used (i.e., in the 2-2 and 3-3 architectures), colors were counterbalanced across participants. The type of trial presented within each block was randomized with the restriction that a color or word could not be repeated more than twice on consecutive trials. A different random order was used for each block and for each participant.

Procedure

Before the test blocks, each participant was given 12 practice trials to help familiarize the participant with the task. In these practice trials, each of the three conditions (congruent, neutral, and incongruent) was represented in the same proportions as they appeared in the actual experiment.

Each trial included a fixation stimulus (consisting of three plus signs), displayed for 700 ms, then a blank screen for 50 ms, followed by presentation of the test stimulus, which remained on the screen until the participant responded. The participant's response triggered a voice-operated relay, after which the experimenter coded the response. The experimenter used one key to signify a correct response. Other keys were used to signify the type of error. The first type of error consisted of false starts, stutters, and other extraneous sounds. The second type of error consisted of reading the word when the task was to name the color: an intrusion error. The third type of error consisted of making an incorrect response that was not an intrusion error. After the response was coded, there was a 1,750-ms intertrial interval before the start of the next trial. The participants received breaks every 24 trials and be-

tween the two blocks. The experimenter remained in the testing room throughout the session to code the participant's vocal responses.

Results

Only response latencies from trials with correct responses were included in the analyses. In addition, response times that fell below 200 ms were assumed to be anticipations, and response times beyond 2,000 ms were assumed to be lapses of attention and were excluded from the analyses. Response times that fell beyond 2.5 standard deviations from each participant's cell mean were also excluded from all analyses. This screening procedure eliminated 1.9% of all correct responses in the 2-2 set size, 2.4% of all correct responses in the 3-3 set size, and 2.5% of all correct responses in the 4-4 set size. Table 6 presents the mean response latencies for both word reading and color naming for each set size.

A 2 (order) × 3 (set size) × 2 (task) × 3 (condition) mixed-factor analysis of variance (ANOVA) was used to examine the data. As expected, there were main effects of task (color naming or word reading), $F(1, 68) = 493.42$, $MSE = 4{,}984$, $p < .0001$; set size (2-2, 3-3, or 4-4 architectures), $F(1, 68) = 7.41$, $MSE = 22{,}707$, $p < .01$; and condition (congruent, incongruent, or neutral), $F(2, 136) = 279.03$, $MSE = 721$, $p < .0001$. In addition, a main effect of order, $F(2, 68) = 5.04$, $MSE = 22{,}707$, $p < .05$, indicated that when word reading was the first task presented, participants averaged 528 ms for all conditions, compared with 561 ms when color naming was the first task. However, it is important to note that order was not involved in any interactions. The analysis did reveal reliable interactions between condition and task, $F(2, 136) = 203.49$, $MSE = 789$, $p < .0001$; between task and set size, $F(2, 68) = 11.33$, $MSE = 4{,}984$, $p < .0001$; and between condition and set size, $F(2, 136) = 5.98$, $MSE = 721$, $p < .001$. Furthermore, there was a highly reliable interaction among condition, task, and set size, $F(2, 136) = 7.46$, $MSE = 789$, $p < .0001$, indicating that not only was the difference between congruency conditions greater in color naming than in word reading but these differences increased as the number of alternatives in the response set increased from two to four. Separate analyses of the word reading and color naming tasks further elucidate these findings.

Word Reading

A 3 (set size) × 3 (condition) mixed-factor ANOVA was conducted on the word reading latencies to determine the effect of set size and congruency on word reading. This analysis yielded only a main effect of condition, $F(2, 138) = 13.39$, $MSE = 90$, $p < .0001$. There was no main effect for set size, $F(2, 69) = 1.71$, $MSE = 9{,}803$, $p > .15$, nor was there a reliable interaction between set size and condition, $F(4, 138) = 0.97$, $MSE = 90$, $p > .40$. Thus, there was no influence of set size on word reading latencies.

Color Naming

Turning to the color naming latencies, the 3 (set size) × 3 (condition) mixed-factor ANOVA yielded highly reliable main effects of both condition, $F(2, 138) = 257.39$, $MSE = 1{,}400$, $p < .0001$, and set size, $F(2, 69) = 10.87$, $MSE = 19{,}146$, $p < .0001$. In addition, the interaction between set size and condition was also reliable, $F(4, 138) = 7.22$, $MSE = 1{,}400$, $p < .0001$, reflecting the fact that the effect of congruency was different as the number of response alternatives increased. When the data were analyzed by using only the neutral and incongruent conditions, the results indicated that as the number of response alternatives increased, the interference effects became larger (80 ms in the 2-2 set size, 142 ms in the 3-3 set size, and 151 ms in the 4-4 set size). A highly reliable interaction between set size and condition, using only the neutral and incongruent conditions, confirmed this observation, $F(2, 69) = 9.58$, $MSE = 1{,}897$, $p < .0002$. In contrast, when the facilitation effect was examined (using only the neutral and congruent conditions), no main effect of congruency was found, $F(2, 69) = 0.93$, $MSE = 617$, $p > .30$, nor was there a reliable interaction between set size and condition using only the neutral and congruent conditions, $F(2, 69) = 1.39$, $MSE = 617$, $p > .25$.

Received January 25, 1996
Revision received December 30, 1996
Accepted January 2, 1997 ■