

Individual Differences in Exemplar-Based Interference During Instructed Category Learning

David C. Noelle

(NOELLE@CNBC.CMU.EDU)

Center for the Neural Basis of Cognition
Carnegie Mellon University
Pittsburgh, PA 15213 USA

Garrison W. Cottrell

(GARY@CS.UCSD.EDU)

Department of Computer Science & Engineering
University of California, San Diego
La Jolla, CA 92093 USA

Abstract

Instructed category learning studies have shown that categorization practice on a fixed set of labeled training exemplars can cause learners to violate explicitly provided categorization instructions. We have previously proposed a connectionist account of this exemplar-based interference effect — an account which predicts that individuals who display initial difficulty in the application of a categorization rule will exhibit greater exemplar-based interference than good rule-followers. In this paper, we report on a study of human instructed category learning performance intended to test this prediction of the model, and we provide the results of additional connectionist simulations which are fit to the human experimental data.

Introduction

Instructed category learning studies have revealed that experience with labeled examples can sometimes cause learners to deviate from previously provided categorization instructions, even when the category labels on the training items are perfectly consistent with the given instructions (Allen and Brooks, 1991; Brooks et al., 1991). Such experiments typically begin with the presentation of an explicit rule for categorizing stimuli. These initial instructions are followed by a sequence of trials in which stimuli are presented to the learner, one at a time. The learner is asked to make a categorization judgment for each stimulus, and this judgment is immediately followed by performance feedback, providing the correct category label for the object. After a substantial period of such training, the learner is presented with novel stimuli and is asked to provide category labels for these novel items. Previous studies have discovered that learners may sometimes violate the instructions that they were given when faced with a novel stimulus, assigning the category label of a similar training set exemplar to the novel item in lieu of accurately applying the given explicit rule.

We have previously presented a connectionist model of instructed category learning which explains this exemplar-based interference effect as emerging from the use of an error-correcting learning rule when learning from examples (Noelle and Cottrell, 1996). This model posits the existence of a working memory network which actively maintains a distributed pattern of activity encoding the explicitly provided categorization rule. The model also includes a categorization network — a system which assigns category labels based on stimulus features. The behavior of the categorization network is modulated by activity in the working memory network, allowing explicit instructions to shape categorization performance. Exemplar-based interference appears when connec-

tion weights in the categorization network are modified, by an error-correcting learning rule, as the result of performance feedback on the training exemplars. This mechanism makes a general prediction concerning exemplar-based interference: difficulty in rule-following should result in larger amounts of interference. If the network exhibits substantial residual error when applying a given explicit categorization rule, this error will produce large weight changes during exemplar-based training, and significant interference will arise. This prediction has a number of corollaries. First, the complexity of the categorization rule should impact interference, with more complex rules producing more interference. Second, a corresponding trend should be seen across individual learners, with individuals who are error prone at rule application exhibiting more interference than those who find rule application easy.

In this paper, we investigate these predictions. We report on a human instructed category learning study which investigates individual differences in exemplar-based interference, and we provide the results of detailed connectionist simulations which model the observed human learning performance.

Individual Differences in Interference

Method

Undergraduate students were asked to learn to categorize a set of simple geometric line drawings into two categories. Each geometric stimulus involved a circle with a radial line, drawn in green on a black computer screen. The circle stimuli varied along two continuous dimensions: size and orientation. Four different sized circles were used, with radii of approximately 5.0 mm, 7.0 mm, 10.0 mm, and 14.0 mm. The number of distinct orientations was also four, with the radial line of the circle rotated counterclockwise from the right-pointing vector by 30°, 60°, 120°, or 150°. Each of the four angles of rotation could be paired with each of the four sizes, producing a set of 16 different stimulus items. These stimuli may be graphically depicted as points in a two dimensional feature space, as shown on the right side of Figure 1. Of the 16 possible circle stimuli, seven were distinguished as training set items. These items are marked with boxes in Figure 1. Four training stimuli were to be placed in one category, called the “black” category here for convenience, and the remaining three were to be placed in the other, called the “white” category.

A typical experimental trial involved the presentation of a circle in the middle of a blank computer screen. Participants were expected to identify the appropriate category for each stimulus, communicating their judgment by depressing the appropriate key on the computer keyboard. No time pressure

was placed on the learners, and accuracy was stressed in initially provided task instructions. Once a category judgment was made for a given stimulus object, a message appeared above the circle indicating if the given classification was correct or not. The correct category label was also explicitly provided at this time. This feedback remained on the display for 2 seconds, after which time the next trial began.

Participants were randomly assigned to one of two experimental conditions: the simple rule condition or the complex rule condition. In each of these two conditions, the experimental session began with the presentation of an explicit rule for categorizing the circle stimuli. The learners in the simple rule condition were told that circles in the “black” category were those of the smallest size, or of the largest size, or rotated 150°. All other circles were to be placed in the “white” category.¹ Participants in the complex rule condition were given similar categorization instructions, only their rule included an “exception” clause. All circles of the smallest size, the largest size, or rotated 150° were to be placed in the “black” category *unless* they were rotated by 60°. All circles with radial lines at 60° were to be placed in the “white” category, even if they were of the largest or smallest sizes. The instructions were designed to ensure that the rules were clearly understood. The four stimuli sizes and the four angles of rotation were graphically displayed on the same screen with the textually presented categorization rule. The rule was described in plain English, making reference to the graphical examples. Furthermore, participants were not allowed to advance to the next stage of the experiment until they demonstrated an accurate memory of the rule by correctly identifying a reworded version of it in a list of three alternatives. Participants also demonstrated retention of the rule by describing it during an informal debriefing following the experiment.

Once categorization instructions were given, learners were presented with 36 blocks of training trials, each block consisting of one presentation of each of the seven training set items, appearing in a random order. Each trial involved the display of one of the stimuli, a categorization judgment on the part of the learner, and a period of performance feedback which provided the correct category label for the object. At the end of this *training phase*, participants were given a short break, during which time they were told that performance feedback would be suspended for the remainder of the experimental session. They were then presented with 8 blocks of trials incorporating all 16 of the possible stimulus objects. The stimuli were presented in a random order, with the learner providing a category label for each object but receiving no feedback concerning the accuracy of such judgments. Following this *testing phase*, the session was paused once more, and learners were given a new categorization rule to apply. They were told that they would soon be asked to classify circles according to the new rule without the benefit of performance feedback. The new rules involved rotating the structure of the original rules in feature space, keeping the complexity of the rules constant. In the simple rule condition, the new rule placed items in the “black” category if they were rotated to 30°, or to 150°, or if they were of the smallest size. The new complex rule was the same, ex-

¹The stimuli used here and the simple categorization rule are derived from the experiments of Nosofsky, Clark, & Shin (1989).

cept all circles of the penultimate size (i.e., 10.0 mm radius) were to be placed in the “white” category, regardless of orientation. Following these new classification instructions, 8 more blocks of trials were given, each involving all 16 stimuli, randomly sequenced. The goal of this final collection of trials was to assess the rule-following ability of each learner when no exemplar-based feedback was made available. This final *rule-following phase* was conducted with new categorization instructions to avoid transfer from the earlier training phase.² The total number of categorization trials experienced by each participant was 508, and these were typically completed within a period of 45 to 55 minutes.

The performance of participants from a third experimental condition is also reported here. The data for these learners were collected during a previous experiment (Noelle et al., 2000). In this third condition, no explicit categorization instructions were given. Participants were asked to learn to categorize the circle stimuli from feedback on the training items alone. These learners experienced 252 training trials with the seven training stimuli, as the instructed learners did, and they were tested, without feedback, on all 16 objects for 128 trials, as before. Since no explicit instructions were given to these participants, no final rule-following test was conducted.

All of the participants in this experiment were undergraduate students enrolled in psychology or cognitive science courses at the University of California, San Diego during the 1996–1997 academic year. They received course credit in exchange for their participation. Data were collected for 36 uninstructed participants, 28 learners in the simple rule condition, and 27 in the complex rule condition. Some students did not appear to be engaged by the task, exhibiting chance level rule-following performance or chance level performance on the training set items even after the 252 training trials. The data for these participants were discarded, leaving valid data for 34 uninstructed learners, 27 learners in the simple rule condition, and 19 learners in the complex rule condition.

Results

The mean frequency of classification responses, averaged over the uninstructed learners, are displayed in Figure 1. The mean results for the instructed participants are shown in Figure 2. The chart on the right side of Figure 1 presents the letter labels which will be used to refer to individual stimulus objects in the discussion, below. The other feature space graphs in these two figures show the frequency with which learners identified objects as being in the “black” category during various phases of the experiment.

Exemplar-based interference, if present, should be found in the testing phase categorization frequencies. Such interference involves a change in categorization performance away from that dictated by the rule instructions and towards that suggested by the distribution of training set items alone. The responses of the uninstructed learners may be taken as a characterization of the category structure suggested solely

²Previous experiments attempted to assess rule-following ability by testing each participant on the original categorization rule, over all 16 stimuli, without feedback, *prior* to the training phase. It was found, however, that such an initial rule-following test, even without performance feedback, impacted performance during the later training and testing phases in a manner which masked interference.

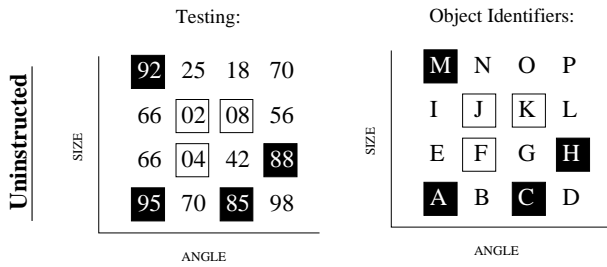


Figure 1: Uninstructed Condition Mean Responses: Categorization results are shown as the percentage frequency with which items were placed in the “black” category. Training set items are marked with boxes, colored according to the assigned category label. Also shown are letter labels for the stimuli, used to reference items in this report.

by the training exemplars. Performance during the final rule-following phase may be used to approximate the accuracy, over all 16 items, with which a learner might have applied the original rule prior to the training phase. Thus, exemplar-based interference may be said to have been present in a given condition to the degree that the pattern of testing phase responding deviated from rule application behavior in the direction of that exhibited by the uninstructed learners.

In order to assess if exemplar-based interference was present in a given experimental condition, careful attention must be given to the amount of error displayed by the participants when they apply an explicit rule without the benefit of exposure to training items. Exemplar-based interference may be said to exist only if deviation from perfect rule application *increases* as a result of performance feedback on the training set. This characterization suggests a quantitative measure of interference involving assessing the deviation from the explicit rule in both the rule-following and testing phases and taking the difference between these two values. Rule-following phase accuracy may be used as an approximation of how well learners might have applied the original explicit categorization rule prior to exemplar-based training. This estimate is compared to categorization accuracy during the testing phase, after training is complete. Using this measure reveals that deviation from the rule did not reliably change in the simple rule condition ($t(26) = -0.664$) but did increase in the complex rule condition, with marginal reliability ($t(18) = 2.02$; $p = 0.06$).

Comparing deviation from the rule across the rule-following and testing phases is not a very powerful test of interference, however. The presence of exemplar-based interference does not entail increased deviation from the rule for *all* stimulus items. Indeed, it is reasonable to expect that the classification performance for some of the stimulus objects will become *more* consistent with the explicit rule as a result of exemplar-based performance feedback, since, for some objects, the category structure suggested by the training items is consistent with that specified by the rule. A more sensitive test of interference would restrict its consideration to those stimulus items for which interference is reasonably expected. A simple operational definition of this expectation can be based upon the uninstructed participant data, expecting

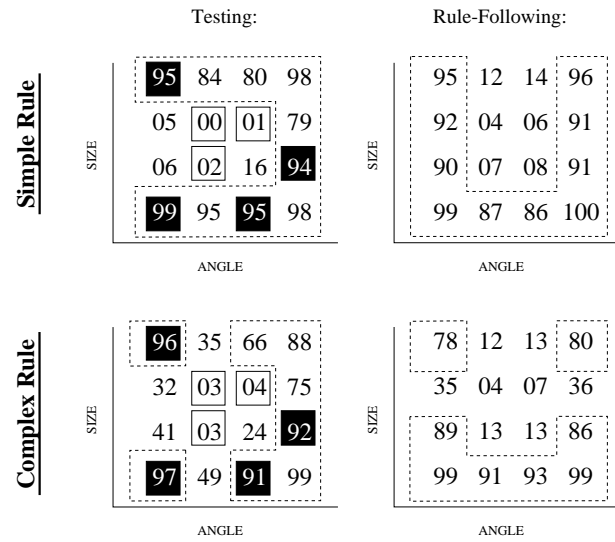


Figure 2: Instructed Conditions Mean Responses: Categorization results are shown as the percentage frequency with which items were placed in the “black” category. Explicitly provided rules are displayed as dashed boundary lines.

interference for those objects which the uninstructed learners, on average, placed in the opposite category as that specified by the explicit rule. The simple categorization rule conflicts with mean uninstructed learner performance at items “E”, “I”, “N”, and “O”. For the complex rule, interference is expected for items “B”, “E”, “I”, and “O”. Restricting our attention to these items, our measure of interference becomes: the increase in mean error, defined as deviation from the rule, from the rule-following phase to the testing phase, averaged only over those items for which interference was expected.³

Making use of this more sensitive measure reveals no reliable interference in the simple rule condition ($M = 0.006$; $SD = 0.244$; $MSE = 0.059$; $F(1, 26) < 1$) but substantial interference in the complex rule condition ($M = 0.220$; $SD = 0.244$; $MSE = 0.060$; $F(1, 18) = 15.478$; $p < 0.001$).

Our previous connectionist model of instructed category learning explained exemplar-based interference as the result of connection weight modifications made during the training phase, driven by an error-correcting learning rule (Noelle and Cottrell, 1996). Under this view, large residual errors will produce large weight changes in the network, producing large amounts of interference. Thus, this model gave rise to the prediction that increased error during rule application (estimated by rule-following phase error) should be accompanied by increased exemplar-based interference. Support for this

³There may be concern that this measure of interference is inappropriate since different rules were used in the rule-following and testing phases. Indeed, it may seem odd to compare categorization error on a specific stimulus (e.g., item “O”) across the two phases when the relationship of that stimulus to the category boundaries changes between the phases. These concerns may be partially alleviated, however, by noting that none of the significance results reported here change if deviation from the rule is averaged over *all* stimuli in the rule-following phase, and this average deviation is then compared to the average testing phase error on stimuli for which interference is expected.

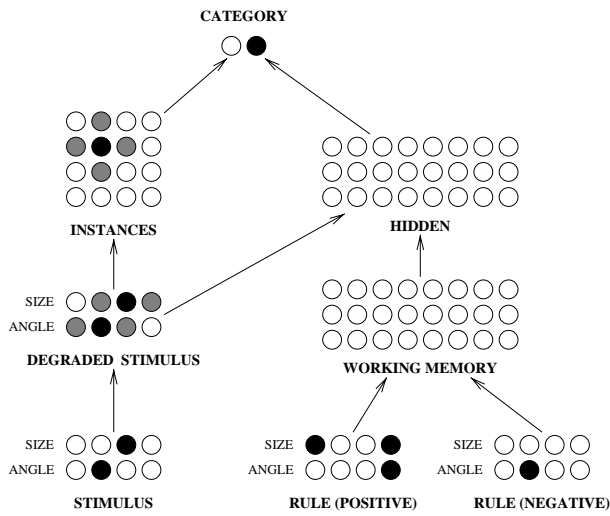


Figure 3: Network Architecture: Each circle represents a standard sigmoidal processing unit, with the units grouped into layers. Arrows represent connections between the units in these layers. Example activations for stimulus “J”, being classified according to the complex rule, are also shown.

prediction may be seen in the difference between the simple rule and the complex rule conditions. The complex rule condition, which elicited a greater degree of rule-following phase error than the simple rule condition (13.5% average error per stimulus object versus 7.9%), elicited a greater degree of interference during the testing phase. This prediction may also be investigated at the level of individual differences. According to the model, poor rule-followers should display more interference than good rule-followers. In order to test this prediction, a correlation was computed over participants in each condition between the mean classification error during the rule-following phase (averaged over all 16 stimuli) and the sensitive measure of interference. A statistically significant positive correlation would verify the prediction. When tested, a marginally reliable *negative* correlation was observed in the simple rule condition ($r = -0.378$; $t(25) = -2.041$; $p = 0.052$), but a robust *positive* correlation was found in the complex rule condition ($r = 0.629$; $t(17) = 3.340$; $p < 0.01$). In brief, the condition which displayed significant interference over all (the complex rule condition) also supported the individual differences prediction of the connectionist model, while the condition which showed no reliable interference (the simple rule condition) revealed a tendency for poor rule-followers to become more consistent with the explicit rule as a result of training.

A New Connectionist Model

Simulation Method

Our connectionist model of instructed learning (Noelle and Cottrell, 1996) has been augmented to provide a detailed account of the experimental results reported here, and new model simulations have been conducted. The network architecture used in these simulations is diagrammed in Figure 3. Standard connectionist processing elements were used,

grouped into layers, the activation level of each processing element being the result of applying a logistic sigmoid to the weighted sum of input activity levels. The activity of each unit was, thus, bounded between zero and one.

The network took a representation of categorization instructions and a pair of stimulus features as input and produced a category judgment at the output layer. The input rule representation included eight units corresponding to the four levels of size and the four levels of angle of rotation used in the human learning experiments. Activating one of these units indicated that all stimuli of the given size or of the given orientation were to be placed in the “black” category. These rule input units are shown in Figure 3 as the “RULE (POSITIVE)” layer. Alongside this layer is a collection of inputs, called “RULE (NEGATIVE)”, which encoded “exceptions” to the positive rule terms. Activating one of these eight “negative” units indicated that all stimuli incorporating the given feature level were to be placed in the “white” category, regardless of the category suggested by the positive terms. As an example, the input representation for the complex rule used in our human learning study is shown in Figure 3.

These explicit rule inputs fed activity to a 24 unit working memory layer. When modeling uninstructed learners, the activity levels of these working memory units were set to zero. The weights on the connections from the rule inputs were bound to be non-negative, forcing the working memory layer to encode explicit rule terms by an increase in the activation levels of its processing elements. These weights were initially set to small random values sampled uniformly from the range $[0.0, 0.4]$. Complete connectivity extended from the working memory layer to a pool of 24 instruction-sensitive hidden units, and these, in turn, provided activity to the category output units. There were no bounds on these weights. The hidden layer also received complete connections from the degraded stimulus layer, which is described below. All of these unrestricted weights were initialized to small random values, sampled uniformly from the range $[-0.5, 0.5]$. Bias values on the working memory and hidden units were initialized to -3.0 in order to encourage sparse internal representations.

Each stimulus was encoded by activating exactly one of the size units and one of the angle units. In order to incorporate perceptual similarity information into the network, this “place coded” stimulus representation was mapped, through connections with fixed weights, to a degraded stimulus representation. In this modified stimulus representation, each unit responded preferentially to a particular stimulus size or stimulus orientation, with partial activity appearing for stimuli of similar sizes or orientations. Levels of partial activation were set to decay exponentially with the number of feature levels separating the given unit from the stimulus. For example, the activity of the “size 2” unit when the stimulus was of “size 4” was set to $e^{-\beta(4-2)}$, where β was a gain parameter which could be modified to fit the model to data. Each stimulus dimension, size and orientation, had its own independent gain parameter, making them analogous to the dimensional attention weights used in models like ALCOVE (Kruschke, 1992).

The “instances” layer contained 16 processing elements, with each unit corresponding to one of the possible stimulus objects. Each unit in this layer received input from exactly one size unit in the degraded stimulus layer and from exactly

one angle unit. Unlike other connections in this network, the activity from these two units was multiplied together, rather than summed, to get the resulting activity of the instances layer unit. Thus, each unit in this layer responded preferentially to a unique stimulus object, and activity declined exponentially with city-block distance in feature space. The weights which gave rise to this pattern of activation were fixed. This representational scheme was adopted because of its success in capturing perceived similarity between stimuli in models such as ALCOVE (Kruschke, 1992). The instances hidden layer provided complete connections to the two category output units, with these weights initialized to small random values uniformly sampled from the range $[-0.5, 0.5]$. The biases on the output units were initialized to -3.0 .

In order to capture human performance, the network had to be able to apply categorization instructions from the very start of the experimental session. Connection weights which allowed the network to produce accurate categorization decisions immediately following explicit instruction were discovered through a training process conducted during a network initialization phase. During this phase, the network was iteratively presented with a randomly sampled categorization rule along with a randomly sampled stimulus object. It was trained to activate the rule-determined category unit for the given stimulus, modifying weights based on squared error at the output layer using the generalized delta rule (Rumelhart et al., 1986). A learning rate of 0.05 was used, with no momentum term. The gain terms used in the degraded stimulus layer were fixed at 0.8 during this initialization training. This phase continued for 5,000,000 training trials, after which the network consistently demonstrated essentially perfect rule application performance. The distribution of stimuli experienced by the network during this initial training was uniform over the 16 items, but the distribution of categorization rules was skewed towards simple category structures. This biased distribution of rules produced a network which exhibited slightly lower residual error when following a simple rule, as compared to a complex one, and it also encouraged the working memory layer to devote more representational resources to the encoding of simple category structures. The skewed rule distribution also reflected a belief that simple rule structures are much more common in the rule-driven categorization experience of most humans.

Once initialized, the network was presented with the same sequence of trials that was presented to the human learners. When uninstructed, the working memory units were turned off and the network was trained on the seven training items for 252 trials. To measure rule-following performance, the appropriate rule was presented at the input, and category outputs were recorded without performance feedback. To measure performance after both instruction and exemplar-based training, the network was given the appropriate rule at its input and trained on the seven exemplars for 252 trials. The network's performance on all 16 stimuli, without feedback, was then recorded. All exemplar-based training was conducted using the generalized delta rule, with a learning rate of 0.5 and no momentum. Only weights from the instances layer to the output category units, and the bias weights on the output units, were modified during this training process.

To simulate limitations in the cognitive resources applied

to the task, random noise was injected into the activation levels of the processing elements in the working memory layer. During each trial, a random deviant was sampled from a zero-mean normal distribution independently for each unit in the working memory layer and the absolute value of this deviant was subtracted from the activation level of the unit. This caused components of the distributed rule representation held in the working memory layer to become weakened. This use of random noise was intended as a simple and abstract way to capture the temporary failure of the working memory system to actively maintain complete representations of categorization instructions. Resampling the noise on every trial was meant to allow for the possibility of refreshing working memory contents from a longer term episodic memory store.

Network simulations were run 50 times for each experimental condition. Each network was initialized with the same set of connection weights, determined during the initialization phase, but both the injected noise and the order of stimulus presentation was randomized for each simulation. The results of each collection of 50 simulations were averaged to produce figures to be compared to mean human categorization behavior. Four free parameters were adjusted to fit the simulations to data. These included the two gain parameters on the exponential decay used in the degraded stimulus layer representation, the gain parameter used on a Luce choice ratio which converted output activity values to probabilities, and the variance of the noise injected into the working memory layer. A simple grid search was conducted over this parameter space to find values for these four parameters which minimized the squared difference in the probability of "black" category assignment between the human learners and the networks, over all experimental conditions.

Simulation Results

The best fit of mean simulation results to the human data was had by sharpening the representation of stimulus orientation slightly (gain of 0.9) over the representation of stimulus size (gain of 0.8). This meant that the angle of rotation was slightly more discriminable by the networks than size. The best fit to the mean data required a Luce choice gain of 2.6 on the output activation levels and working memory injected noise with a variance of 0.3. The resulting probabilities of "black" category membership, as predicted by the model, are shown in Figure 4. That diagram also displays the variance accounted for by the model, over stimulus items, for each condition. Notice that, like the humans, the network simulations exhibit no interference in the simple rule case on average (a value of -0.017 in the interference measure previously used with the human data), but they show substantial interference when given the complex rule (0.251).

These simulations involved only a single source of individual variation: the ability to actively maintain an accurate representation of the categorization instructions. Individual differences in exemplar-based interference, then, were to be explained in terms of the weight modifications which were driven by the rule application error introduced by a failure to maintain explicit rules in working memory. Variable working memory ability was reflected by the noise variance parameter in these networks. Thus, in order to examine individual differences in interference, networks with a range of values

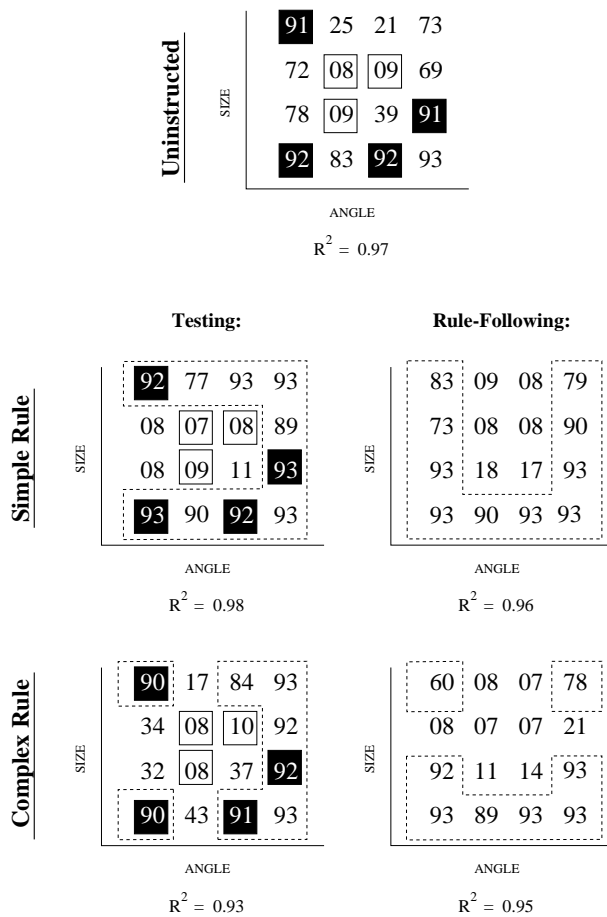


Figure 4: Simulation Results: Categorization results are shown as the mean probability with which items were assigned to the “black” category. Training set items are marked with boxes, colored according to the assigned category label. Variance accounted for is listed for each condition separately.

for this noise parameter had to be compared. Given a collection of networks with a variety of noise levels, the correlation between rule-following error and interference may be measured. Recall that human learners displayed a positive correlation when confronted with the complex rule, but showed a marginally negative correlation when given the simple rule.

When the variance on the noise injected into the working memory units was sampled uniformly from a bound range, these simulations matched the human findings. For example, if noise variance was sampled uniformly from the set $\{0.0, 0.1, 0.2, 0.3\}$, then the correlation between rule-following error and interference was reliably negative for the simple rule ($r = -0.283$; $t(198) = -4.15$; $p < 0.0001$) and reliably positive for the complex rule ($r = 0.412$; $t(198) = 6.37$; $p < 0.0001$). Similar results were found when the noise parameter was sampled in a manner sensitive to the observed distribution of human rule-following performance. This sampling was done by finding individual network simulations that matched, as closely as possible, the rule-following phase accuracies exhibited by individual human learners. These best match networks were found by varying the noise variance be-

tween 0.0 and 0.4. When correlations between rule-following error and interference were calculated for such participant-matched samples of network simulations, a positive correlation was found for the complex rule case ($r = 0.545$; $t(17) = 2.68$; $p < 0.05$), and no correlation was found in the simple rule case ($r = -0.166$; $t(25) = -0.840$).

Conclusions

The magnitude of exemplar-based interference was found to be sensitive to the complexity of the explicitly provided categorization instructions, with more complex categorization rules producing more interference. Also, in situations which elicit robust interference, a reliable correlation across individuals is observed: increased error at explicit rule application is paired with increased exemplar-based interference. A connectionist account of these effects, in which interference arises as the result of an error-correcting learning process, was found to fit the human performance data fairly closely.

Acknowledgements

This work was supported, in part, by the NIH through a National Research Service Award (# 1 F32 MH11957-01) from the National Institute of Mental Health, awarded to the first author. We extend our thanks to Craig R. M. McKenzie, James L. McClelland, David Plaut, and the members of the UCSD-based *Gary's Unbelievable Research Unit* and the CMU based *PDP Research Group*, as well as to five anonymous reviewers, for their comments on this work.

References

- Allen, S. W. and Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, 120(1):3–19.
- Brooks, L. R., Norman, G. R., and Allen, S. W. (1991). The role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, 120(3):278–287.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44.
- Noelle, D. C. and Cottrell, G. W. (1996). Modeling interference effects in instructed category learning. In Cottrell, G. W., editor, *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, pages 475–480, La Jolla. Lawrence Erlbaum.
- Noelle, D. C., Cottrell, G. W., and McKenzie, C. R. M. (2000). Modeling individual differences in the specialization of an explicit rule. (in preparation).
- Nosofsky, R. M., Clark, S. E., and Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(2):282–304.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, chapter 8, pages 318–362. MIT Press, Cambridge, Massachusetts.