On the Normativity of Failing to Recall Valid Advice

David C. Noelle (NOELLE@CNBC.CMU.EDU)
Center for the Neural Basis of Cognition; Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

Instructed category learning tasks involve the acquisition of a categorization skill from two sources of information: explicit rules provided by a knowledgeable teacher and experience with a collection of labeled examples. Studies of human performance on such tasks have shown that practice categorizing a collection of training examples can actually interfere with the proper application of explicitly provided rules to novel items. In this paper, the normativity of such exemplar-based interference is assessed by confronting a model of optimal memory performance with such a task and comparing the "rational" model's behavior with that exhibited by human learners. When augmented with a rehearsal mechanism, this optimal memory model is shown to match human responding, producing exemplar-based interference by relying on memories of similar training set exemplars to categorize a novel item, in favor of recalling rule instructions.

Introduction

Contemporary studies of human category learning have tended to focus on the acquisition of general knowledge about a new concept exclusively from exposure to a collection of labeled examples. In common learning environments, however, students attempting to learn a categorization skill are frequently provided with more than a set of training examples. In particular, learners are often explicitly instructed in the nature of a new category before being presented with instances. They are provided with definitional sentences and explicit rules (e.g., "an equilateral triangle has at least two sides of the same length" or "bugs with six legs are insects"). Direct instruction of this kind can rapidly provide a basic understanding of a new category, while experience with examples can further shape and refine that initial understanding (Klahr and Simon, 1999).

While it is common for the process of explicit instruction following and the process of induction from examples to cooperate to produce quick and robust learning, there are situations in which these two learning processes actually compete. Specifically, practice at classifying a set of training examples can cause learners to violate explicitly provided categorization rules when classifying novel items. Extensive experience with examples can lead learners to categorize novel instances according to similarity to training items, rather than according to categorization rules communicated through explicit instruction. Thus, novel items which are highly similar to training examples from another category come to be misclassified as a result of practice.

This exemplar-based interference effect, in which experience with examples interferes with proper instruction following, was investigated by Allen and Brooks (1991), as well as others (Brooks et al., 1991; Neal et al., 1995; Noelle and Cottrell, 2000). Such interference in category learning is mirrored by similar difficulties in a wide variety of learning contexts, such as when students come to solve math or science problems by analogy to previously seen problems, rather than by application of formal principles and techniques communicated through direct instruction. Learners appear to have a tendency to disregard perfectly valid explicit advice in favor of knowledge induced from experiences with examples.

Exemplar-based interference might be seen as the result of limitations of the cognitive system, such as imperfect working memory efficacy (Noelle and Cottrell, 2000) or difficulties recalling and applying abstract, linguistically encoded, rules. There is another alternative, however. It is possible that human learners neglect explicit instructions in favor of experienced exemplar-similarity information because the latter form of information tends to be more reliable in a wide variety of learning contexts. Exemplar-based interference may be the result of an essentially normative process of weighting sources of category information according to the previously established utilities of those sources.

There are many aspects of common learning situations which may encourage students to rely more heavily on examples than on explicit rules. Consider, for example, how the instructions provided by teachers are frequently approximate and heuristic. Advice is often implicitly limited to a particular range of circumstances, and there are often exceptions, even within this range, to explicitly provided rules. Also, teachers are sometimes in error. In short, human learners may have strong reasons to doubt the perfect accuracy of offered categorization rules. In comparison, exemplar similarity may be seen as a highly reliable indicator of category membership. Most categories, after all, involve clusters of similar objects, suggesting that similarity might be the best tool for predicting the category labels of novel instances.

Even if considerations of teacher reliability are ignored, there are other rational reasons for a learner to

rely preferentially on training experiences. In general, recalling past experiences with features similar to those of the current situation is often more useful than recalling dissimilar experiences. Thus, when faced with the task of categorizing a novel stimulus item, learners may be naturally inclined to recall other similar items rather than an explicit rule, which, due to its linguistic encoding, may bear little surface similarity to the situation at hand. Also, the recollection of experiences which are recent and frequently recurring is, on average, more useful when facing a novel challenge than recalling rare experiences from one's distant past. Thus, when performing an instructed category learning task, it may be reasonable for a learner to selectively recall the training items which were recently and repeatedly studied in favor of a briefly presented rule. In short, we may conjecture that exemplar-based interference arises from a rational tendency to rely on similar, recent, and frequent past experiences when faced with a novel situation.

In order to evaluate this conjecture, this paper reports on the modeling of the exemplar-based interference results of Allen and Brooks (1991) using the normative, or "rational", account of memory formulated by Anderson (1990). The goal is to investigate the degree to which exemplar-based interference can be explained in terms of a Bayes optimal learning process, given some assumptions about the common demands placed on human memory. The human performance results are reviewed first, followed by a description of Anderson's optimal memory model. The results of applying the model to this domain are then presented.

Human Performance

Allen and Brooks (1991) performed a number of experiments demonstrating the way in which experience with labeled training exemplars can interfere with instructed rule following. In their Experiment 1, learners were asked to categorize cartoon illustrations of fictional animals into one of two categories, based on how the animals were said to construct their homes: the "builders" and the "diggers". The appropriate category for each animal was strictly determined by it's physical features. Each animal was composed of specific selections for five binary attributes: angular body shape or rounded body shape, spots or no spots, short legs or long legs, short neck or long neck, and two legs or four legs. Only three of these attributes were ever relevant for classification, however: body shape, presence or absence of spots, and leg length. The animals were always depicted against color backgrounds, displaying four different outdoor environments. From this space of $2^5 \times 4 = 128$ different possible stimuli, only 16 were actually used. These 16 items were carefully chosen to include two animals with each possible level of the three relevant attributes. The irrelevant features were selected so that each stimulus item would have exactly one "partner" item — an item which differed from it only in the presence or absence of spots. Otherwise, each animal differed from each other animal in at least two attributes.

Experimental participants were provided with explicit categorization rules for discerning the "builders" from the "diggers". These always took the form of "2 of 3" rules, in which a target category was described as all animals with at least two of a list of three features (e.g., builders have two or more of the following features: angular body shape, spots, long legs). The rules were carefully chosen so that the 16 stimuli were equally split between the two categories. Also, the exemplars were partitioned into a training set and a testing set so that no two "partnered" items were in the same set. This resulted in exactly half of the testing set items having their partner items in the opposite category. These testing items were the ones for which interference was predicted.

The learners were presented with a training phase which consisted of seeing each of the 8 training set items five times, presented in a random order, for a total of 40 trials. When a stimulus image appeared on the screen, learners were to categorize it as quickly as possible, without sacrificing accuracy. Then, a sequence of two slides would be shown, illustrating how the animal actually constructed its home, identifying it as a builder or a digger. A subsequent testing phase involved soliciting categorization responses from the participants without providing any form of feedback on their decisions. During this testing phase, each training set stimulus was presented 4 times and each testing set stimulus was presented once, for a total of 40 testing trials.

There were two main results of this experiment. First, accuracy on the items whose "partners" were in the opposite category was much worse than on the other testing set items — around 55% correct as compared to 80%. This was a strong indication of exemplar-based interference. Second, the response time for correctly classified items was much larger for items whose "partners" were in the opposite category. This was interpreted as extra caution on the part of the learners when facing these "tricky" stimulus items. In other words, even when exemplar-based interference did not cause error, it at least caused a slowing of behavior.

Allen and Brooks argued that explicit memories for individual stimulus items played an important role in the production of this interference effect. The presentation of a testing set stimulus was seen as provoking a recollection of that item's "partner" in the training set, with the category label of that training set item often being assigned to the new stimulus in lieu of a label based on explicit rule application. Following this intuition concerning the centrality of memory to this effect, we have attempted to model these data using a previously explicated account of optimal memory performance.

Anderson's Rational Memory

The hypothesis explored here is that the behavior of the learners examined by Allen and Brooks can be characterized as normative — as the natural result of employing a memory system which is optimal in a Bayesian sense. This raises the question of how an optimal memory system would respond in this domain. Anderson and Mil-

son (1989) have proposed a "rational" model of memory which might be employed to address this question.

Initially, one may think that an optimal memory is a perfect memory. *Everything* is to be stored in every detail, without degradation, for an unlimited amount of time. This overlooks one very important function of memory, however, and that is to recall only those memories which are relevant to the current task. Without this ability of selective recall, a memory is essentially useless, even if (or especially if) it contains every detail that was ever experienced. Thus, the task faced by an optimal memory is the identification of those memory traces which would be most useful in the current situation.

In Bayesian terms, the goal is to determine, for each memory trace, the probability that that trace would be useful in the current situation. In Anderson's model, this is called the "need probability" of a trace. An optimal memory is seen as one which retrieves exactly those traces with the highest need probabilities in the current context. The question then becomes one of calculating the need probability for each memory trace.

In this model, the need probability is seen as a function of two components: the *desirability* of the trace and the *association* between the trace and the current context. The desirability of a memory trace is a measure of the average utility of the trace — a kind of base rate of appropriateness. The desirability of a trace is to be induced from its history of use. Recent and frequent retrieval of a memory trace is indicative of high desirability. The association between the trace and the current context is a kind of normalized likelihood of the context given that the trace is needed. This term increases the need probability with increased similarity between the context and the trace. Both of these components of the need probability are seen as normative properties of the situation, unbiased by predispositions of the agent. In brief, the optimal memory system computes the need probability of each memory trace, conditioned on the current context and on the history of past retrievals of that trace.

Mathematically, if A represents the event that a given memory trace is needed in the current context, H_A represents the complete retrieval history of that trace, and Q is the current context, then the conditional need probability is $P(A|H_A \ \& \ Q)$, which may be decomposed as follows:

$$P(A|H_A \ \& \ Q) \ = \ P(A|H_A) \times \frac{P(Q|A)}{P(Q)}$$

Note that this assumes that Q and H_A are both independent and conditionally independent with respect to A. If Q is taken to be composed of a collection of mutually independent features, then this expression may be written as:

$$P(A|H_A \ \& \ Q) \ = \ P(A|H_A) \times \prod_{i \in Q} \frac{P(i|A)}{P(i)}$$

This formulation allows for the separate calculation of a *history factor*, $P(A|H_A)$, and a *context factor* which measures the association between the memory trace and each feature of the current context, $P(i|A)$.

The calculation of the history factor requires some assumptions about the desirability of memory traces. Each trace is taken to start at some desirability level, λ_0, when it is first generated. Over the range of memory traces, these initial desirabilities are assumed to have a gamma distribution with parameter b and index ν. This means that no traces have an initial desirability of zero, most have some small initial desirability, and a very few have a high value for this variable. Furthermore, desirability is assumed to decay exponentially over time, with a decay rate of δ, where this rate of decay varies over the traces. It is assumed that δ is exponentially distributed with parameter α. Together, these assumptions paint a picture of memory traces with various initial desirabilities, decaying exponentially over time at various rates. Some memory traces start out with a high desirability and decay only slowly, like, say, the trace for your own name. Other traces start out with a low probability of use, like instructions on how to help a heart attack victim, but the desirability does not decay much with time. Some memories are very important but only for a short time, such as the memory for how much money was handed to a cashier before receiving change. Most trivia start out with a low desirability and decay rapidly.

One phenomenon not captured by this characterization is the way in which certain memory traces might become very useful again, after a long period of unimportance. To remedy this oversight, it is assumed that memory traces occasionally experience "revivals", at which time their desirabilities are returned to their original levels. The probability of a revival of a memory trace is assumed to decay exponentially with the time since the trace's introduction, with rate β.

This formulation provides a characterization of the probability distribution of possible trajectories of desirability over time. Recall, however, that what is needed is the distribution of histories of actual trace retrievals:

$$P(A|H_A) \ = \ \frac{P(A \ \& \ H_A)}{P(H_A)}$$

If we assume that a trace is retrieved with a probability proportional to its desirability, we can compute $P(H_A)$ by integrating over all possible values of initial desirability, decay rate, and revival history. This value is:

$$P(H_A) \ = \ \int\int \ P(H_A|\delta \ \& \ R) \ p(\delta) \ p(R) \ \ d\delta \, dR$$

where δ is a decay rate and R is a particular revival history. Note that, in this expression, the initial desirability has already been integrated over. The main term in this double integration has the form:

$$P(H_A|\delta \ \& \ R) \ = \ \frac{b^{\nu}(n+\nu-1)!}{D^{n+\nu-1}(\nu-1)!} \ \prod_{i=1}^{n} e^{-\delta(H_i - r_i)}$$

where n is the number of retrievals in H_A, H_i is the time of the *ith* retrieval, r_i is the time of the revival which most

immediately preceded the *ith* retrieval, and D is:

$$D = b + \frac{1}{\delta} \sum_{j=0}^{m} \left(1 - e^{-\delta(R_{j+1} - R_j)}\right)$$

where m is the number of revivals, and R_j is the time of the *jth* revival. All other variables in these expressions are parameters from the previously discussed probability distributions. In short, an expression for the value of $P(H_A)$ is available in the form of the double integral above.[1] This double integral ranges over an infinite space of δ values and possible revival histories. In order to estimate the value of this expression, a Monte Carlo integration may be performed, sampling decay rates and revival histories from their respective distributions. In this way, an estimate of $P(H_A)$ can be calculated.

Note that $P(A \ \& \ H_A)$ can be calculated in exactly the same fashion as $P(H_A)$ simple by including an additional retrieval of the memory trace at the current moment. As previously noted, the ratio of these two probabilities is the needed history factor, $P(A|H_A)$.

The calculation of the context factor is much easier to perform, mostly due to some simplifying assumptions. To compute the contribution of the association between the trace and the current context, it is assumed that the trace is composed of features which contribute independently to the need probability of the trace. These features are assumed to be mutually independent, even when conditioned on any feature of the current context. Thus, the context factor can be written as:

$$\prod_{i \in Q} \frac{P(i|A)}{P(i)} = \prod_{i \in Q} \frac{P(A|i)}{P(A)} = \prod_{i \in Q} \prod_{x \in A} \frac{P(x|i)}{P(x)}$$

All that remains is to determine the associative strengths between features of the current context and features of the memory trace, expressed as $P(x|i)$, which may be selected in a manner sensitive to the specific stimuli used.

Anderson and Milson (1989) showed that this optimal memory model matched human performance in many ways. This calculation of the probability of retrieval was found to predict recency and frequency effects, and the model was shown to be consistent with effects arising from varying the temporal spacing between the presentations of stimuli. This complex retrieval probability computation accounted for effects of word frequency on the memorization of word lists, priming effects, and various fan effects. Most all of these calculations were performed with fixed values for the distribution parameters: $b = 100$, $\nu = 2$, $\alpha = 2.5$, and $\beta = 0.04$.

Modeling Exemplar-Based Interference

Following the theorizing of Allen and Brooks (1991), their instructed category learning task can be viewed as a memory task. When initially given the explicit rule for categorizing the fictional animals, the learner must remember this rule, and it must be recalled when it is needed to categorize a stimulus item. The rule need not always be recalled, however, as it will be sufficient in many cases to simply remember a previous presentation of the specific stimulus being viewed and its corresponding category label. This characterization of the task makes Anderson's rational memory model applicable to an optimality analysis of instructed category learning.

A computer program was written which simulated the performance of Anderson's rational memory on the experimental task examined by Allen and Brooks (1991). Initial instruction involved the creation of a memory trace for the given categorization rule, and the retrieval of that trace for ten consecutive time steps, representing a study period. After this instruction period, the training set items were presented to the optimal memory, one at a time, in the same manner as they were presented to human participants. With each presentation, the need probability of each existing memory trace was estimated in the context of the current stimulus. The memory trace with the highest need probability among those traces that contained a category label was retrieved from the memory.[2] The category label of the retrieved memory trace was taken to be the response provided by the optimal memory system to the current stimulus. Note that the memory trace for the explicit rule was seen as containing the correct category label for every stimulus item.

During the training phase, the solicitation of a categorization judgment from the memory was followed by the incorporation of performance feedback information. The memory system responded to feedback by immediately retrieving the memory trace corresponding to the current stimulus, or, if this was the first presentation of the given item, by generating and retrieving a new trace for the stimulus, marked with the given category label.

After the training phase, the optimal memory experienced a testing phase equivalent to that presented to the human learners, involving a mix of training set items and new testing set items. The protocol for memory trace retrieval during the testing phase was the same as during training, except that none of the newly generated memory traces contained category label information, as no feedback was provided to the humans during this phase. Categorization errors made by the memory system during the testing phase were examined for signs of exemplar-based interference: relatively poor accuracy on those testing set stimuli whose "partner" items in the training set were in the opposing category.

To calculate the history factor of the need probabilities, the same parameters that were used by Anderson and Milson (1989) were used in this simulation: $b = 100$,

[1] Note that this expression is different than that provided in the appendix of Anderson and Milson (1989). When this error was brought to the attention of the authors, they provided the software that they had used to perform their calculations. It was discovered that the error was only in their appendix and not in their software.

[2] During the testing phase it was possible that the memory trace with the highest need probability would be a memory of a previous presentation of an unlabeled item. Such a memory would not be of much use for making a categorization judgment. Thus, this retrieval was restricted only to those memory traces which contained explicit category information.

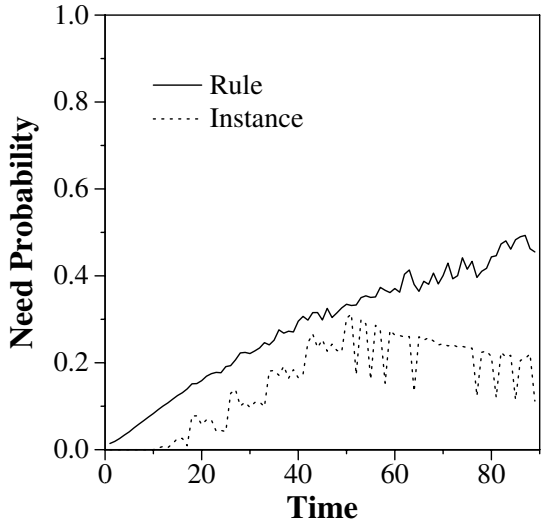

Figure 1: Results from the Optimal Memory Model: The need probability of the rule memory trace is plotted against the maximum need probability among the traces for the training set items. Note that the training phase ran from time step 11 through 50, and the testing phase ran from time 51 through 90.

$\nu = 2$, $\alpha = 2.5$, and $\beta = 0.04$. To calculate the context factor, the presentation of a stimulus was seen as providing a context consisting of 5 binary features (i.e., the attributes of the fictional animals) and one 4-ary feature (i.e., the background). Memory traces were seen as containing these six features, plus an optional category label. The associational strength between context and trace features was taken to be $P(x|i) = 0.65$. These features were taken to be pictorial in nature, so the memory trace for the explicit verbal rule contained none of these features. The Monte Carlo integration process employed by the optimal memory model consistently used $100,000$ samples in the calculation of each need probability estimate.

A summary of the results of this computation are shown in Figure 1. Plotted in that graph is the calculated need probability of the explicit rule memory trace and the highest need probability over the training set exemplar traces, both over time. Note that the training phase began at time step 11 and ended at time step 50, and the testing phase ran from time step 51 through time step 90. The primary result shown in this graph is that the rule always dominated over the exemplars. This meant that the rule was always retrieved in preference to traces for previously viewed items. In other words, the optimal memory produced perfect rule following behavior with no sign of interference. Even when the optimal memory system was modified to stochastically retrieve traces in a manner proportional to their need probabilities (rather than always retrieving the trace with the highest need probability), errors on stimulus items with "partners" in the opposite category averaged only 12%, as compared to the 45% error exhibited by humans.

These results were found, however, to be very sensitive to the associational strength that was used, $P(x|i)$. If this value was substantially increased above 0.65, then the memories for the training set items would immediately and persistently dominate over the trace for the rule. Under such higher settings of the associational strength, the optimal memory model would produce interference during the appropriate portions of the testing phase, but it would not produce expected behavior early in the session. In particular, the explicit rule would almost never be used. In short, this initial simulation of the optimal memory model of instructed category learning did not match human performance very well at all.

Anderson had some similar problems with his rational memory model when he compared its performance to human behavior (Anderson, 1990). While human responding matched his rational memory calculations in a number of domains, there were some aspects of human performance which could only be fit by the model with the help of an additional assumption. This assumption was that the system would covertly rehearse recently retrieved traces. He added to the memory model a rehearsal buffer which contained the 4 most recently retrieved memory traces. On each time step, each trace in the rehearsal buffer had a 0.2 probability of being rehearsed on that time step. Rehearsal simply involved the retrieval of that trace from memory. Increasing the number of retrievals of a trace through rehearsal would expand its retrieval history, H_A, and would thereby increase the history factor, $P(A|H_A)$, for that trace. Anderson added this rehearsal strategy, admitting that it stepped beyond the bounds of an optimality analysis. Still, such an augmented analysis was considered worthwhile, since it could show that human performance is optimal up to the inclusion of such rehearsal strategies. Indeed, that was exactly what Anderson demonstrated for a number of memory phenomena.

Following Anderson's lead, the optimal instructed category learning simulation was augmented with a 4 element rehearsal buffer. As in Anderson's work, the probability of rehearsal for each item in the buffer was set to 0.2 per time step. The memory trace for the instructed rule was allowed to occupy the buffer and be rehearsed, just like any other memory trace. The associational strength parameter was kept at 0.65.

Adding this rehearsal mechanism had a substantial impact on the behavior of the optimal memory, as shown in Figure 2. With rehearsal, the explicit rule maintained its perceived utility through much of the training phase, but was overcome by exemplar similarity by the time the testing items were presented. This produced consistent errors on those stimuli whose "partners" were in the opposite category. When traces were retrieved stochastically, in proportion to their need probabilities, the frequency of error on such items was 42%, comparing favorably to the 45% error exhibited by human learners. Thus, the rational memory model, when augmented with rehearsal, appears to be consistent with the observed in-

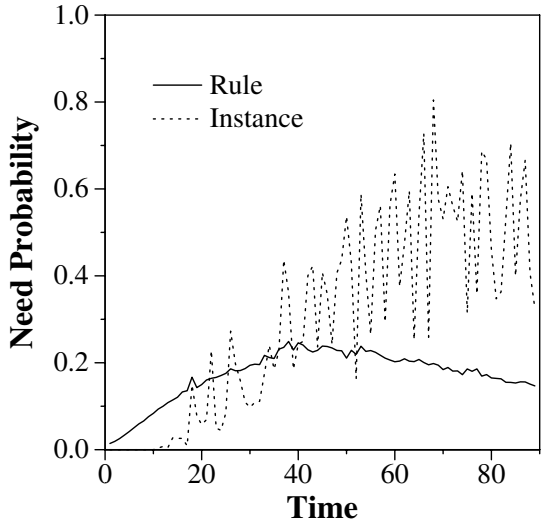

Figure 2: Results from the Optimal Memory Model With Rehearsal: Once again, the need probability of the rule memory trace is plotted against the maximum need probability among the traces for the training set items.

terference effect in instructed category learning.

Discussion

In many situations, it is more useful to remember a highly similar episode from the past than to recall generally applicable instructions. The rational memory model of Anderson and Milson (1989) is a formalization of the process of optimally predicting when such a situation has arisen. The unaugmented optimal memory model specifies that, within the experimental design of Allen and Brooks (1991), the explicit rule should almost always be preferred if similarity is not very predictive (i.e., when the associational strength is low), and a memory for specific instances should almost always be preferred if similarity is sufficiently predictive (i.e., when the associational strength is high). This is not consistent with human performance, however, where errors on "tricky" testing set items appeared only 45% of the time.

However, if the rational memory model is augmented with a rehearsal mechanism, as is needed to explain performance on other memory tasks (Anderson, 1990), the resulting need probabilities match human performance much more accurately. This suggests that the interference effect of interest may arise in the interaction between an optimal memory mechanism and a rehearsal strategy. One prediction of this calculation is that experimental manipulations which hinder rehearsal will reduce exemplar-based interference.

Note that, in these simulations, the memory trace for the explicit rule shared no features with the stimulus presentation contexts. This was intended to model the fact that the stimuli were pictorial, while the rule was linguistic. In fact, if the features itemized in the explicit rule are associated with the corresponding stimulus features with the same associational strength as used elsewhere in these simulations (0.65), the explicit rule comes to dominate over exemplar memory traces, even in the augmented model. It is a surprising fact is that this property of the model actually reflects human responding. Exemplar-based interference virtually disappeared when Allen and Brooks (1991) presented the animal stimuli not as pictures but as word lists — allowing the stimulus features and the explicit rule terms to literally match.

In summary, while this analysis does not rule out other potential explanations of exemplar-based interference, it offers the tantalizing possibility that the human tendency to ignore explicit instructions in favor of information provided by example experiences may be essentially adaptive when considered within the context of the common demands placed on the cognitive systems responsible for learning and memory.

Acknowledgements

This work was supported, in part, by a National Research Service Award (# 1 F32 MH11957-01) from the USA National Institute of Mental Health. Thanks are extended to Garrison W. Cottrell, Richard Anderson, and two anonymous reviewers for their helpful comments.

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.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Studies in Cognition. Lawrence Erlbaum, Hillsdale, New Jersey.

Anderson, J. R. and Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96(4):703–719.

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.

Klahr, D. and Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, 125(5):524–543.

Neal, A., Hesketh, B., and Andrews, S. (1995). Instance-based categorization: Automatic versus intentional forms of retrieval. *Memory & Cognition*, 23(2):227–242.

Noelle, D. C. and Cottrell, G. W. (2000). Individual differences in exemplar-based interference during instructed category learning. In Gleitman, L. R. and Joshi, A. K., editors, *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, pages 358–363, Philadelphia. Lawrence Erlbaum.