Extensions to the delta rule for associative learning

John K. Kruschke and Amy L. Bradley
Indiana University

June 12, 1995

The delta rule of associative learning has recently been used in several models of human category learning, and applied to categories with different relative frequencies, or base rates. Previous research has emphasized predictions of the delta rule after extensive learning. Our first experiment measures the relative acquisition rates of categories with different base rates, and the delta rule significantly and systematically deviates from the human data. We suggest that two additional mechanisms are involved, namely, short-term memory and strategic guessing. Two additional experiments highlight the effects of these mechanisms. The mechanisms are formalized and combined with the delta rule, and provide good fits to the data from all three experiments.

Several recent models of category learning in humans incorporated the *delta rule* of associative learning (Rumelhart, Hinton, & Williams, 1986). The delta rule posits that the growth in the strength of association between a cue and an outcome is error driven: The associative strength changes in magnitude proportionally to the discrepancy, or error, between the actual magnitude of the outcome and the magnitude predicted by the current associative strength.

Gluck and Bower (1988) demonstrated that a form of apparent base rate neglect observed in human learners could be produced by a simple instantiation of the delta rule. Gluck and Bower (1988) used a simulated medical diagnosis paradigm, in which participants learned which diseases (categorical outcomes) were associated with which symptoms (cues). Apparent base rate neglect is the tendency to diagnose a particular symptom as a *rare* disease when the symptom's normative, Bayesian diagnosticity is *equal* for a common and the rare disease. The effect, and the model's ability to reproduce it, have been replicated several times (Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Lewandowsky, 1995; Nosofsky, Kruschke, & McKinley, 1992; Shanks, 1990). Kruschke (1995) has shown, however, that the model cannot account generally for base rate neglect, and that the delta-rule model must be extended to include rapid attentional shifts across cues. The delta rule also has been used in the ALCOVE model of category learning (Kruschke, 1992, 1993b) to build associations between memory exemplars and categories, and the model can account for a number of phenomena in human category learning (Choi, McDaniel, & Busemeyer, 1993; Kruschke, 1993a; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Nosofsky & Kruschke, 1992).

Because of the pioneering work of Gluck and Bower (1988) on category base rates, much of the recent research involving applications of the delta rule to human category learning has used differential category base rates (e.g. Estes et al., 1989; Kruschke, 1995; Lewandowsky, 1995; Markman, 1989; Nosofsky et al., 1992; Shanks, 1990, 1991a, 1991b, 1992). Much of the emphasis has been placed on predictions of the model after asymptotic learning. A model of learning, however, should also address the course of learning. In particular, because of the concentration of research involving differential base rates, the model should address the relative learning rates of categories that occur with different relative frequencies. Past research has not addressed this issue in detail. Recently, Kruschke and Erickson (1994) found that ALCOVE could not quantitatively reproduce the learning curves of four exemplars that occurred with relative frequencies of four, three, two and one. The model did not learn the high frequency exemplars rapidly enough in the early learning trials, and it could not learn the low-frequency exemplars quickly enough in the intermediate trials.

The goal of the research presented in this article is to investigate the inability of the delta rule to match quantitatively the relative acquisition rates of categories that occur with different relative frequencies. The first experiment measures relative acquisition rates in human learners for simple single-cue, single-outcome associations. It is shown that the

This research was supported by an NIMH FIRST Award to the first author. The second author participated in the research in partial fulfillment of the requirements for the Bachelor of Arts with Honors in Psychology at Indiana University. For their valuable comments on this research, the authors thank Michael Erickson, Michael Fragassi, Gabriel Frommer, Brian Gygi, Eliot Hearst, Michael Kalish, Armando Machado, and Linda Smith. For helping to conduct the experiments, the authors thank Abbey Clawson, Sarah Countryman, Matthew Durkee, Sharon (Michelle) Estes, Toby Matherly, Beth Okeon, Laura Owen, Debbie Reas, Amy Stierwalt and Daniel Vote.

Correspondence concerning this article should be addressed to John K. Kruschke, Dept. of Psychology, 1101 E. 10th St., Indiana University, Bloomington IN 47405–7007. E-mail: kruschke@indiana.edu, Web page: http://www.indiana.edu/~kruschke/. This article is available at http://www.indiana.edu/~kruschke/deltarule_abstract.html.

delta rule deviates from the human data significantly and systematically. We claim that to account for the data, the delta rule must be extended with at least two additional mechanisms: short-term memory and strategic guessing. Two additional experiments are presented to demonstrate robustly the effects of short-term memory and strategic guessing in this learning paradigm. Finally, the notions of short-term memory and strategic guessing are formalized and combined with the delta rule, and the extended model is shown to provide good fits to the data.

Experiment 1: Frequency Effects

In Experiment 1 we measured learning as a function of training frequency, for simple, deterministic associations of single symptoms with single diseases. Each of six symptoms was a perfect predictor of one of six diseases. One symptom-disease pair occurred with a relative frequency of six, another occurred with a relative frequency of five, another with a relative frequency of four, and so on, down to a relative frequency of one. The goal was to obtain quantitative data to which the delta rule could be fit, from a minimalist paradigm in which stimuli consisted of only single symptoms and the mapping from symptoms to diseases was deterministic, unlike the multiple-cue, probabilistic designs used in much previous research.

Method

Participants. Forty-six volunteers participated for partial credit in an introductory psychology course at Indiana University.

Apparatus. Participants sat in front of a PC-type computer located within individual, sound dampened, dimly lit, ventilated cubicles. Responses were entered by pressing keys on the standard keyboard. The experiments were programmed using Micro Experimental Laboratory (MEL) (Schneider, 1988). The same apparatus was used for all experiments reported in this article.

Design and procedure. Each of six single symptoms was a perfect predictor of one of six diseases. Out of a total of 210 trials, one symptom-disease pair occurred 10 times, another occurred 20 times, another 30 times, and so on, up to the most frequent pair occurring 60 times. The 210 trials were randomly permuted so each participant experienced a different sequence.

At the beginning of the experiment, participants read instructions on the computer screen at their own pace. The instructions did not explicitly specify that there was a one-to-one correspondence of symptoms and diseases, nor did they explicitly indicate that the symptoms were perfect (non-probabilistic) predictors.

For each participant, six symptom labels were randomly chosen from a list of 22 possible labels (ear aches, skin rash, back pain, dizziness, sore muscles, stuffy nose, nausea, hair loss, blurred vision, twitching, fever, wheezing, insomnia, flatulence, perspiration, fatigue, coughing, palpitation, halitosis, anemia, ennui, incontinence). Disease labels were single letters D, F, G, H, J, and K, assigned to symptoms randomly for each participant.

On each trial, a symptom was displayed near the center of the screen, with a response prompt below it reading, "Diagnose as one of D, F, G, H, J, or K." The participant indicated his or her diagnosis by pressing the corresponding letter key on the keyboard. If the participant did not respond within 30 sec, a tone sounded with the word "FASTER" on the screen. If the participant did respond, the word "CORRECT" or "WRONG" appeared on the screen, as appropriate. In any case, the computer then displayed the phrases, "This patient has disease D/F/G/H/J/K. After you have studied this case (up to 30 seconds), press the space bar to see the next one." If the participant studied the case for more than 30 sec, a tone sounded with the phrase "You have only 30 seconds to study each case" on the screen, and the next case appeared automatically. Unlimited rests were provided after trials 80 and 160. At those points the screen displayed the message, "You may now rest a few seconds. Press the space bar to continue." The experiment lasted about 35 min.

Results and Discussion

Figure 1 plots the proportion correct as a function of five-trial blocks, with a separate curve for each relative frequency (1–6). Any given block includes data from five consecutive trials, in which any single participant saw only a subset of the diseases. Collapsed across 46 participants, however, even the low-frequency diseases occurred at least several times in every block. In order better to reveal the gradual course of acquisition and the differences between the learning curves, blocks with relatively few trials had to be used, but the blocks could not contain too few trials or else the low frequency diseases would have occurred too rarely in each block for statistical purposes. As a reasonable compromise, we chose a block size of five trials. Data are shown only for the first half of training, because the latter half of training shows only near-perfect performance on all six diseases.

As one might expect, performance on the more frequent diseases improves more rapidly than for the relatively infrequent diseases. Collapsing across 105 trials, the overall proportion correct increased with the relative frequencies, with values of 0.768, 0.830, 0.845, 0.865, 0.926, and 0.922 (sic), respectively. Overall proportion correct on the highest frequency disease was reliably better than that on the lowest frequency disease, $\chi^2_{(1)} = 52.01$, $p < 0.0001$. This pattern of results is the same as that found by Kruschke and Erickson (1994), where performance on the high-frequency cases rose extremely rapidly in the very early trials, and performance on the low-frequency cases rose fairly rapidly soon thereafter.

Fit by the Delta-Rule Model

For purposes of exposition, we envision the delta rule to be instantiated in a connectionist network, with one input node per symptom, and one output node per disease (cf. Gluck & Bower, 1988; Rumelhart et al., 1986). When symptom *s* is

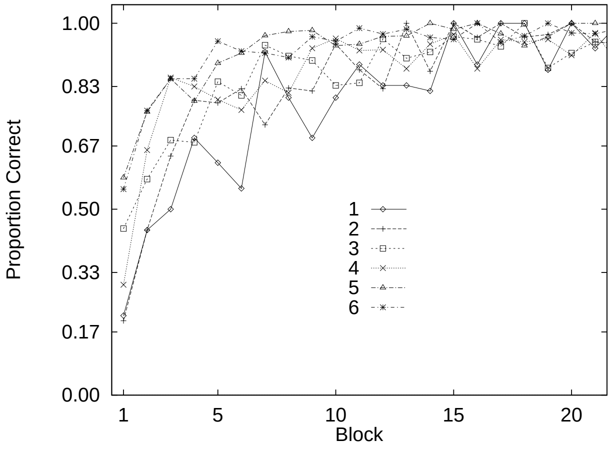

Figure 1. Results of Experiment 1. Numerical labels in the legend indicate the relative frequency of the symptom-disease pair.

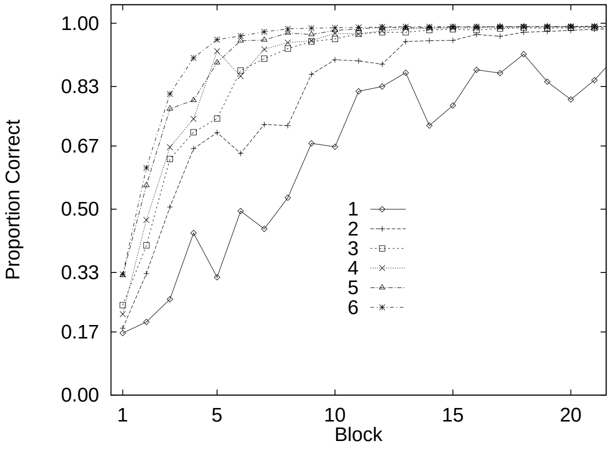

Figure 2. Best fit of the basic delta-rule model to the results of Experiment 1. Numerical labels in the legend indicate the relative frequency of the symptom-disease pair.

present, the corresponding input node has activation $a_s = 1$; otherwise the input node has zero activation. Activation from the symptoms can spread to the disease nodes via weighted connections. The activation strength, a_d, of disease d is given by the sum of the association weights, w_{ds}, times symptom activations, a_s:

$$a_d = \sum_s w_{ds} a_s. \tag{1}$$

The activation of category node d reflects the strength with which the stimulus is thought to be a member of that disease category. The probability of actually choosing disease D for the diagnosis is given by a variant of the Luce (1963) choice rule,

$$p_D = \exp(\phi a_D) \Big/ \sum_d \exp(\phi a_d) \tag{2}$$

where ϕ is a scaling constant. In previous research, ϕ was a freely estimated parameter. In the applications reported in this article, however, ϕ is set at a fixed value that yields $p_D = 0.99$ (an arbitrarily chosen limiting value) when $a_D = 1.0$ and $a_d = 0.0$ for all $d \neq D$. This is done in order to isolate and contrast the other mechanisms of the models. When there are six candidate diseases, $\phi = 6.39$.

After corrective feedback is supplied, the association weights are adjusted using the delta rule, so that the change in weight is given by:

$$\Delta w_{ds} = \lambda(t_d - a_d)a_s \tag{3}$$

where λ is a learning rate, and t_d is the *teacher* for disease d, given by[1]

$$t_d = \begin{cases} 1 & \text{if } d \text{ is the correct category} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The model was trained on the same 41 sequences experienced by the participants. The model was fit to the 254 frequencies in a three way table generated by crossing 21 blocks with 6 disease frequencies and 2 outcomes (correct or wrong). Because the model was trained on the same sequences as the human participants, the marginal frequency of a particular disease in a particular block is fixed; hence, there are 126 degrees of freedom in the data. The basic delta-rule model consumes one degree of freedom with its single free parameter. The discrepancy of the data from the model was measured by the log-likelihood statistic $G^2 = 2\sum_i f_i \ln(f_i/\widehat{f_i})$, where f_i is the observed frequency in cell i, $\widehat{f_i}$ is the predicted frequency in cell i, and the sum is taken over all cells in the table (Wickens, 1989). The best fit was determined by a hill-climbing parameter search.

Figure 2 shows the best fitting predictions of the basic delta-rule model to the results of Experiment 1. The best fitting learning rate was $\lambda = 0.214$, yielding $G^2_{(125)} = 363.49$, far exceeding the critical χ^2 of 155.93 for a Type I error rate of .01. The model has two systematic deficiencies. First, it cannot raise performance on the high-frequency diseases quickly enough. For example, in the first block, the model predicts the proportion correct for the highest-frequency disease to be 0.32, whereas human learners achieved 0.57. A second deficiency lies in extended poor performance on the rare diseases over several blocks.[2]

A careful examination of the training sequences suggests

[1] Because of the particular training patterns used in these experiments, identical results are obtained if the teacher values are "humble,"

$$t_d = \begin{cases} \tilde{1} = \max(1, a_d) & \text{if } d \text{ is correct} \\ \tilde{0} = \min(0, a_d) & \text{otherwise,} \end{cases} \tag{4}$$

wherein the activation of a node can exceed its target value (either less than 0.0 or greater than 1.0) without penalty (Kruschke, 1992).

[2] When ϕ is allowed to vary freely, the best fitting parameter values to the results of Experiment 1 are $\phi = 4.61$ and $\lambda = 0.465$, yielding $G^2_{(124)} = 188.34$, still far exceeding the critical χ^2 of 154.8 for a Type I error rate of .01. The deficiencies of the model pointed out in the main text persist.

why human learners might be able to achieve such rapid increases in performance. For the high-frequency diseases, it is not uncommon for cases to occur repeatedly on several consecutive trials. If participants were using short term memory, they could easily respond correctly to successive repetitions of a symptom after seeing the correct diagnosis in the first trial of the string of repetitions. By contrast, the delta rule has no form of short-term memory. For the lowest-frequency disease, on the other hand, we discovered that participants performed far above chance (1/6) even on their very first case of that disease. If participants assume that there is a one-to-one correspondence of symptoms with diseases, then they might achieve such performance by strategically guessing: If they encounter a symptom they do not yet know, then they should choose a disease they have not yet been shown. The basic delta rule makes no use of strategic guessing. The next two experiments are designed to highlight those mechanisms in human performance.

Experiment 2: Massed vs. Distributed Learning

Experiment 2 is intended to demonstrate the involvement of short-term memory (STM) in simple associative learning. As in Experiment 1, participants had to learn six symptom-disease pairs, but in Experiment 2 the relative frequencies were equal and the repetition of pairs on successive trials was directly manipulated. In blocks of 12 trials, three of the pairs always occurred in successive trials, and the other three pairs never occurred in successive trials. This arrangement is analogous to a comparison of *massed* and *distributed* practice, respectively, in recall tasks. We expected that performance on the second occurrence of the massed cases would be excellent, because learners would have just seen the correct diagnosis on the previous trial and STM would enable accurate responding to the same case on the subsequent trial. Performance on the second occurrence of the distributed diseases should not be as high, however, because STM would not endure as strongly from the previous occurrence several trials before.

An ancillary prediction was that we might observe a classic *spacing effect* (Ebbinghaus, 1885) for the first occurrence in each block. That is, performance on the first occurrence of distributed cases might be better than performance on the first occurrence of the massed cases.

Method

Participants. Thirty volunteers participated for partial credit in an introductory psychology course at Indiana University.

Design and procedure. Six symptom-disease pairs occurred twice in every block of 12 trials. Three of the pairs were massed, so that the two occurrences appeared in consecutive trials, and the other three pairs were randomly distributed, so that they never occurred on consecutive trials (neither within nor across blocks). Distributed cases were further constrained so that they never appeared immediately before and after two massed cases. The massed cases were separated by two distributed trials, and the massed cases always appeared in the same order in each block, so that they would be maximally separated. The first massed case could occur on trial 1, 2 or 3, determined randomly for each subject.

An example should clarify these details. Let the three massed pairs be denoted M1, M2 and M3, and the three distributed pairs be denoted d1, d2 and d3. For two consecutive blocks the sequence could be as follows: d1, M1, M1, d3, d2, M2, M2, d3, d1, M3, M3, d2, d1, M1, M1, d2, d3, M2, M2, d2, d1, M3, M3, d3.

The instructions were identical to Experiment 1. The only change in procedure was that no rest breaks were given, because the study time on each trial provided opportunities for rest, and because there were fewer trials overall. The experiment lasted about 25 min.

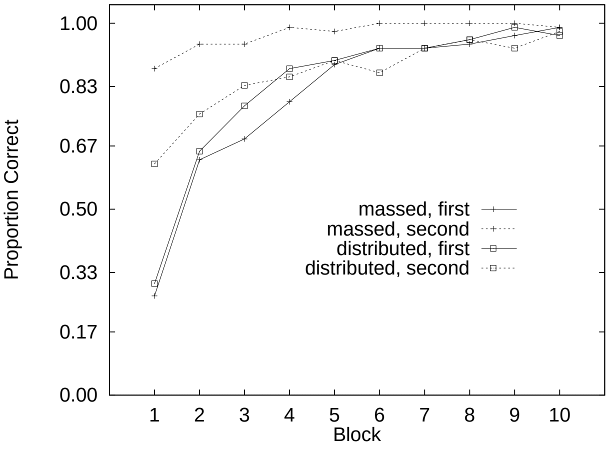

Figure 3. Results of Experiment 2.

Results and Discussion

Figure 3 plots the proportion correct for each occurrence of the massed or distributed cases as a function of training block. Performance on the second occurrence of the massed cases was nearly perfect, and reliably better than performance on the second occurrence of the distributed cases (collapsed across the 10 blocks, $\chi^2_{(1)} = 70.70$, $p < 0.0001$). We interpret this outcome to arise from the selective benefit of STM for the massed cases. We also found a trend toward an effect analogous to the classic spacing effect, with performance on the first occurrence of the distributed cases being somewhat (but not reliably) better than performance on the first occurrence of the massed cases (collapsing across 10 blocks, proportion correct of .829 vs. .803, $\chi^2_{(1)} = 1.96$, n.s.).

Fit by the Delta-Rule Model

The model was trained on the same 30 sequences experienced by the participants. The model was fit to the 80 frequencies in a four way table generated by crossing 10 blocks

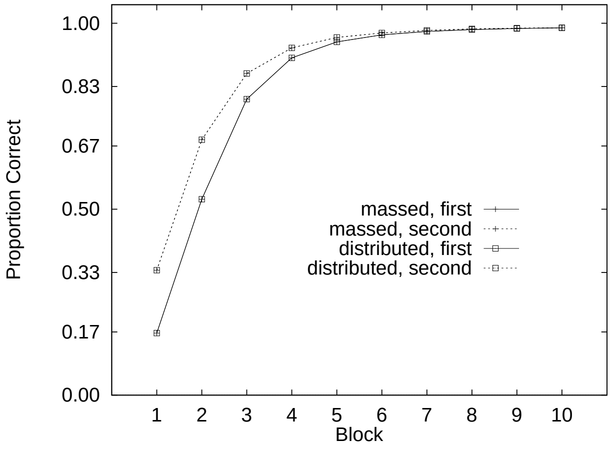

Figure 4. Best fit of the basic delta-rule model to the results of Experiment 2.

with 2 spacings (massed or distributed) and 2 occurrences (first or second) and 2 outcomes (correct or wrong). The marginal frequencies of each spacing and occurrence were fixed by the design; hence, there are only 40 degrees of freedom in the data. The basic delta-rule model uses one free parameter, yielding 39 degrees of freedom for the log-likelihood fit statistic.

Figure 4 shows the best fitting predictions of the basic delta-rule model. The best fitting parameter value was $\lambda = 0.145$, yielding $G^2_{(39)} = 349.41$, far exceeding the critical χ^2 of 62.43 for a Type I error rate of .01. The model cannot show any effect of spacing; it predicts identical performance for the massed and distributed cases.[3]

Experiment 3: Phased Learning

Experiment 2 demonstrated robust effects of massed vs. distributed training, which we interpreted in that context as the effect of STM. The delta-rule model could not account for that effect. In explaining the results of Experiment 1 we hypothesized a second mechanism, namely, a guessing strategy based on an assumption that there was a one-to-one correspondence of symptoms to diseases. Experiment 3 is designed to provide additional evidence for the existence of that strategy.

Experiment 3 retained the same relative frequencies of the symptom-disease pairs as Experiment 1, but distributed their occurrence such that diseases were phased in, over the course of training. The first phase of 10 trials consisted of repeated practice on one symptom-disease pair, then the next phase of 20 trials included 10 occurrences of the initial pair plus 10 occurrences of a second pair, randomly intermixed. The third phase consisted of 30 trials, and included 10 occurrences each of the first two pairs, plus 10 occurrences of a new third pair (randomly intermixed), and so on, until the last phase included 60 trials, 10 trials of each of six symptom-disease pairs. If learners apply a response strategy as suggested previously, then performance on the first occurrence of the symptom-disease pairs in the later phases should be well above chance (1/6). Indeed, perfect application of the strategy would yield perfect performance on the first occurrence of the disease introduced in the sixth phase.

Method

Participants. Forty-one volunteers participated for partial credit in an introductory psychology course at Indiana University.

Design and procedure. The first phase of training consisted of 10 trials of symptom-disease pair 1. The second phase consisted of 10 trials of pair 1 plus 10 trials of pair 2, randomly interspersed. The third phase consisted of 10 trials each of pairs 1 and 2, plus 10 trials of pair 3, randomly intermixed, and so on, through six phases. There were a total of 210 trials. Phase boundaries were not marked in any way; the participant experienced a continuous stream of 210 trials. There were no marked rest breaks. The procedure was identical to Experiments 1 and 2. In particular, the instructions were identical, so participants had no explicit indication that the six cases would be phased in across training. The experiment lasted about 30 min.

Results and Discussion

As expected, performance on the first occurrence of a symptom-disease pair was higher for later phases, as shown in Figure 5. For the sixth pair, performance on the first occurrence was nearly perfect (proportion correct = 0.93). Binomial tests indicate that performance was above chance (1/6) for the fourth-, fifth- and sixth-phase pairs (the critical proportion correct is 0.281, for a null hypothesis of $p = 1/6$, $N = 41$, and one-tailed Type I error rate of .01). We interpret this result as indicating that participants used a guessing strategy: Upon encountering a symptom they did not know, they tended to select one of the diseases they had not yet seen.

After the third occurrence of each disease, performance was nearly perfect throughout the remainder of training, even when new symptom-disease pairs were introduced in later phases. Therefore, those data are not shown in Figure 5, and are not fit in subsequent model fits.

The data also show that performance on the first trial (first occurrence of the first pair) was significantly worse than chance (only 1 of 41 responses was correct). This outcome may be merely an unusual sample from an underlying population with true probability at chance (1/6), or it might indicate that participants had a systematic prior bias to press particular keys more than others, and the pseudo-random assignment of disease labels (keys) to diseases happened not

[3] When ϕ is allowed to vary freely, the best fitting parameter values are $\phi = 4.53$ and $\lambda = 0.311$, yielding $G^2_{(38)} = 252.29$, still far exceeding the critical χ^2 of 61.16 for a Type I error rate of .01. The fundamental deficiency persists: The model cannot show any effect of spacing and predicts identical performance for the massed and distributed cases.

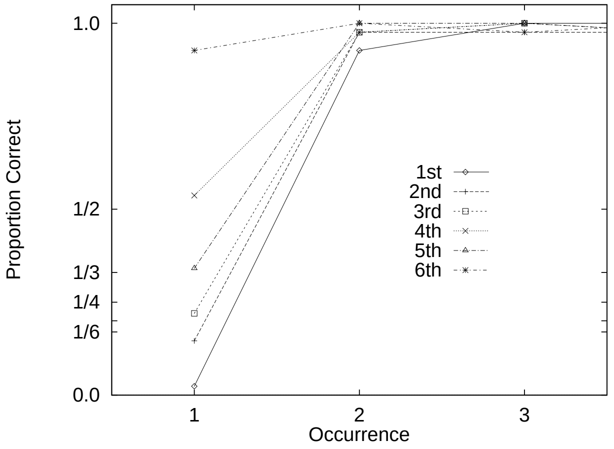

Figure 5. Results of Experiment 3. The legend indicates the phase of learning.

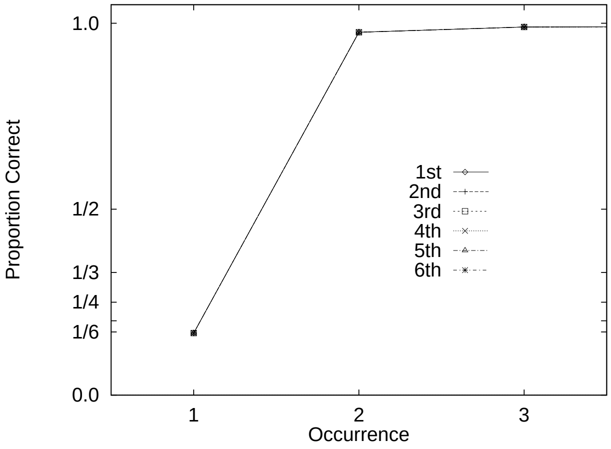

Figure 6. Best fit of the basic delta-rule model to the results of Experiment 3. The legend indicates the phase of learning. All six curves are superimposed

to average out that bias. Finally, the data show an inversion in the anticipated rise in proportion correct with phase, in that participants performed somewhat, but not significantly, better on the first occurrence of the fourth disease than of the fifth disease (.537 vs. .341, $\chi^2_{(1)} = 3.17$, n.s.).

Fit by the Delta-Rule Model

The model was trained on the same 41 sequences experienced by the participants. The model was fit to the 36 frequencies in a three way table generated by crossing 3 occurrences with 6 phases and 2 outcomes (correct or wrong). The marginal frequencies of each phase and occurrence were fixed by the design; hence, there are only 18 degrees of freedom in the data. The basic delta-rule model uses one free parameter, yielding 17 degrees of freedom for the log-likelihood fit statistic.

Figure 6 shows the best fitting predictions of the basic delta-rule model. The best fitting parameter value was $\lambda = 0.835$, yielding $G^2_{(17)} = 173.33$, far exceeding the critical χ^2 of 33.41 for a Type I error rate of .01. The model cannot show any effect of learning phase; it predicts chance performance (1/6) for the first occurrence of every phase's new disease.[4]

An Extended Model

Three experiments have shown the inadequacy of the delta-rule model to account for simple association learning in humans. We suggested that humans use at least two mechanisms not reflected in the delta-rule model, namely, short term memory and strategic guessing. Experiments 2 and 3 were designed to show robust empirical evidence for those mechanisms at work. In the remainder of the article we describe a candidate formalization of those mechanisms, and its fit to the data from the three experiments.

The goal is not to argue for the adequacy of this particular formalization, but to argue that (1) adding mechanisms of short term memory and strategic guessing to the basic delta rule is important for modeling human learning, and that (2) insofar as the extended model better fits the data, mechanisms of STM and strategic guessing are indeed at work in human performance.

Short Term Memory

Short term memory (STM) is implemented simply as another set of association weights v_{ds} between symptom s and disease d. When a symptom-disease pair occurs, the corresponding STM weight is set to 1.0. The STM weight decreases as a function of number of trials since last occurrence, the assumption being that successive trials retroactively interfere with STM.

In order to quantify that decreasing function, we appealed to the classic data reported by Waugh and Norman (1965). In their experiment, participants heard a list of 16 single-digit numbers, the last of which had occurred only once previously in the list. The participant's task was to recall the number in the list that immediately followed the previous occurrence of the last number. There were two important independent variables: (1) the number of interfering items between the last number and its previous occurrence, and (2) the temporal rate at which the lists were presented. Waugh and Norman (1965) found that the temporal rate had no reliable effect on recall performance, and concluded that the dominant cause of forgetting in STM is retroactive interference and not time-based, spontaneous decay. For our purposes, we are more interested in the effect of the number of interfering items. Waugh and Norman (1965) instructed their participants to rehearse only the digit just presented, and not to rehearse any

[4] When ϕ is allowed to vary freely, the best fitting parameter values are $\phi = 6.49$ and $\lambda = 0.822$, yielding $G^2_{(16)} = 173.27$, far exceeding the critical χ^2 of 32.0 for a Type I error rate of .01. The same deficiency persists.

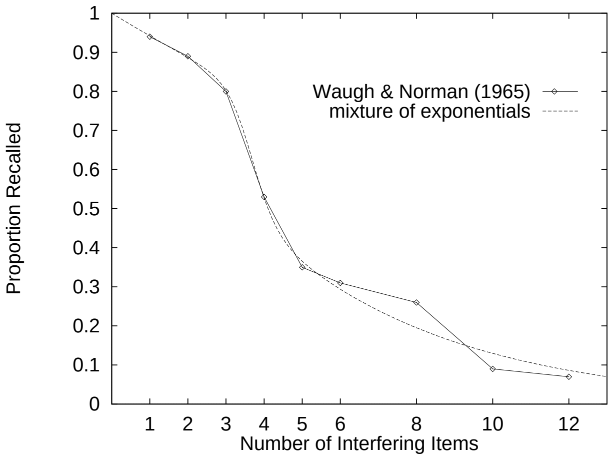

Figure 7. Data from Waugh and Norman (1965) and best fit by a mixture of exponentials.

digits previously presented. It could be assumed, therefore, that there was no rehearsal of successive *pairs* of numbers, so that recall of the number succeeding the probe was based purely on STM. The proportion of correct recall, corrected for guessing, appears in Figure 7 (adapted from Figure 1C, p.91, of Waugh and Norman (1965)).

Visual examination of the data suggests that they can be described as a slow exponential drop for a few trials, followed by a rapid decline to a lower exponential tail. We therefore fit the data with a sigmoidal mixture of exponentials, so that the STM strength, $v_{ds}(n)$, as a function of the number of interfering trials, n, is given by

$$\begin{aligned} v_{ds}(n) &= [1\text{–}\mathrm{sig}(n,\gamma,\omega)] \exp(-n/\sigma_1) \\ &\quad + \mathrm{sig}(n,\gamma,\omega) \exp(-n/\sigma_2) \end{aligned} \tag{6}$$

where σ_1, σ_2, γ, and ω, are freely estimated constants, and where the mixing coefficient $\mathrm{sig}(n,\gamma,\omega)$ is the sigmoidal function

$$\mathrm{sig}(n,\gamma,\omega) = 1.0 - 1.0/(1.0 + \exp[-\gamma(n-\omega)]). \tag{7}$$

The constant, ω, specifies the trial at which the curve will drop rapidly from the high level to the low level. The constant, γ, controls the steepness of the drop. The constant, σ_1, determines the height of the early trials, prior to the drop, and the constant, σ_2, determines the height of the later trials, after the drop.

The best fitting curve is shown in Figure 7, and has parameter values $\gamma = 3.14$, $\omega = 3.64$, $\sigma_1 = 16.8$ and $\sigma_2 = 4.90$. Visual inspection suggests that the fit is excellent. If we assume that all 400 observations contributing to each datum were independent, then the model can be rejected, $G^2_{(5)} = 18.55$, $p < .005$. Assuming smaller values for the number of effectively independent observations per datum yields acceptable values of G^2.

Our goal in fitting the data from Waugh and Norman (1965) is merely to obtain some reasonable curve describing the reduction of STM strength as a function of the number of interfering trials. This exercise in curve fitting is mute regarding the underlying mechanisms of interference. The best fitting curve is henceforth fixed, and incorporated directly into the extended model of category learning described below. Thus, the value of an STM weight is computed using Equation 6, with n set to zero whenever the corresponding symptom-disease pair occurs, and otherwise incremented on every trial. The utilization of the STM weights in diagnosis and learning is described later.

Strategic Guessing

We have provided evidence that participants use a guessing strategy for symptoms they do not yet know, so that they tend to choose diseases they do not yet know. To formalize this idea, we must formalize what it is not to know a symptom, and we must formalize how to choose previously unlearned diseases.

Lack of knowledge about a stimulus can be formalized as the degree of uncertainty in diagnosis. Uncertainty can be measured as the *normalized entropy,* S, of the probability distribution across the candidate diseases:

$$S = \sum_d^N p_d \log p_d \Big/ \log(1/N) \tag{8}$$

where p_d is the probability of disease d, and N is the number of candidate diseases ($N = 6$ in the experiments reported here). The normalized entropy is 1.0 when the diseases are all equally probable, and is 0.0 when when one disease has probability 1.0 and the others have probability 0.0.

To the extent that there is uncertainty in the diagnosis, we want to suppress previously learned diseases. A straightforward means by which to suppress previously learned diseases is to apply *negative* activation to all symptoms absent from the stimulus, and propagate that inhibition via the associations to the disease nodes. Those diseases with strong (positive) association weights from absent symptoms will be inhibited. Those diseases not yet learned, that is, with zero or small association weights, will not be suppressed because the inhibition is not propagated to them. We therefore set the symptom activations as follows:

$$a_s = \begin{cases} 1 & \text{if symptom } s \text{ is present} \\ \mu S & \text{if symptom } s \text{ is absent} \end{cases} \tag{9}$$

where $\mu < 0$ is a freely-estimated constant representing the maximum possible inhibition applied to the symptom nodes, and S is the normalized entropy defined previously in Equation 8.

Combining STM and Strategic Guessing with the Delta Rule

With STM and strategic guessing formalized, we must now specify how they interact with each other and with the delta rule. We will describe these interactions by explaining the sequence of processing during a single trial.

When a stimulus is presented at the beginning of a trial, the input nodes corresponding to presented symptoms are activated ($a_s = 1$ for s present), and the other input nodes remain inactive ($a_s = 0$ for s absent).

The presentation of a stimulus also causes interference in STM, so that the STM weights decrease. This is implemented by incrementing the trial counters of all the STM weights, and setting new STM values according to Equation 6.

Activation is then propagated to the disease nodes, using both the long-term and short-term associations. The total activation, b_d, of disease node b, is given by

$$b_d = \sum_s w_{ds} a_s + \sigma \left[\sum_s v_{ds} a_s - \sum_s w_{ds} a_s \sum_s v_{ds} a_s \right] \qquad (10)$$

where w_{ds} is the long-term association weight from symptom s to disease d, v_{ds} is the short-term weight, and σ is a freely estimated mixing constant. Equation 10 is motivated by considering the long-term and short-term activations to be loosely analogous to probabilities of recall from long-term and short-term memories, respectively, and considering the mixed probability to be the probability of recall from either LTM or STM [$p(\text{LTM or STM}) = p(\text{LTM}) + p(\text{STM}) - p(\text{LTM})\, p(\text{STM})$; cf. Waugh and Norman, 1965, p.93].

The probability of choosing disease D is then determined from the total activations, and is computed analogously to Equation 2,

$$p_D = \exp(\phi b_D) \Big/ \sum_d \exp(\phi b_d) \qquad (11)$$

where ϕ is a scaling constant. As described previously, ϕ is set at the fixed value (6.39 for $N = 6$) that yields $p_D = 0.99$ when $b_D = 1.0$ and $b_d = 0.0$ for all $d \neq D$.

With the disease probabilities determined, the uncertainty of the diagnosis is measured by computing the normalized entropy, using Equation 8. For example, when all diseases have equal probability (1/6), the normalized entropy, S, is 1.0. At another example, when one disease has probability 0.99 and the others have probability 0.002, the normalized entropy, S, is 0.04.

The guessing strategy is then applied, to the extent that there is uncertainty from the initial diagnosis. This is done by applying negative activation to the input nodes of the absent symptoms, according to Equation 9, and then propagating activation once again to the disease nodes using Equation 10. Choice probabilities for each disease are re-computed using Equation 11. It is these probabilities, which combine both STM and the guessing strategy, that are taken as the models' predictions for the trial.

At this point in the trial, corrective feedback is supplied, and the short-term and long-term weights are adjusted. For short-term weights, connections from any present symptom to the correct disease are set to 1.0. In the experiments reported here, every trial has only one symptom, consequently only one short-term weight is set to 1.0.

Long-term weights w_{ds} are adjusted on the basis of error computed *only from long-term weights*. That is, weights are adjusted using Equation 3, wherein a_d is computed from Equation 1. Thus, STM affects only the response probabilities in the short term, but does not affect long-term learning.

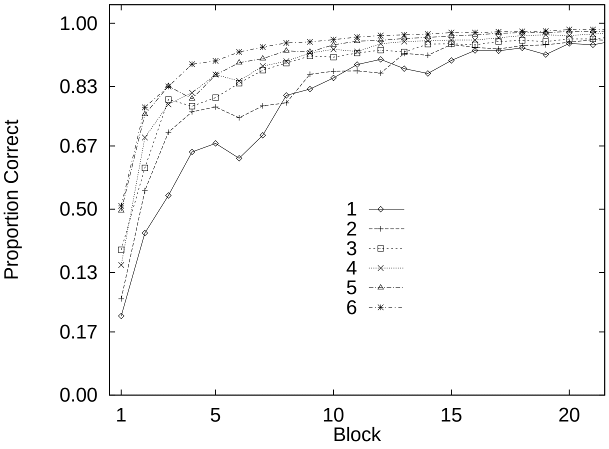

Figure 8. Best fit of the extended model to the results of Experiment 1. Numbers in the legend indicate the relative frequency of the symptom-disease pair.

Various theories of human STM might suggest modification of this assumption. For example, STM might be used for rehearsal and strengthening of corresponding LTM associations. On the other hand, STM might have the opposite effect: STM improves performance, thereby reducing error, consequently impeding learning that is exclusively error-driven as in the delta rule. This is analogous to the inattention hypothesis for explaining poorer learning in massed trials (e.g., Hintzman, 1976): In massed trials, STM can be used to reduce error and hence reduce the "attention" devoted to that trial. For our present purposes, however, we will assume that STM and strategic guessing are used only for diagnosis, and do not affect learning. This is consistent with a major point of this article, that the delta rule might be retained, intact, as long as other mechanisms are included for STM and guessing strategies.

In summary, the extended model retains the learning rate λ from the basic delta-rule model, and adds two parameters. The influence of STM in diagnosis is governed by the mixture coefficient σ (Equation 10). In particular, when $\sigma = 0$, there is no effect of STM. Application of strategic guessing involves one free parameter, μ, which specifies the maximal amount of inhibition applied to absent symptoms (Equation 9). In particular, when $\mu = 0$, there is no effect of the guessing strategy. When both $\sigma = 0$ and $\mu = 0$, the extended model becomes the basic delta-rule model.

Fit to Frequency Effects

The best fit of the extended model to the data from Experiment 1 is shown in Figure 8. Best fitting parameter values were $\lambda = 0.0503$, $\mu = -0.699$, and $\sigma = 0.526$, yielding $G^2_{(123)} = 115.76$. This is an excellent fit to the data, and the model cannot be rejected statistically. In the extended model,

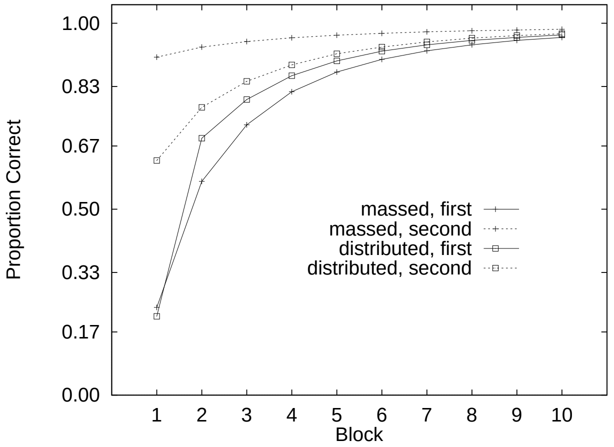

Figure 9. Best fit of the extended model to the results of Experiment 2.

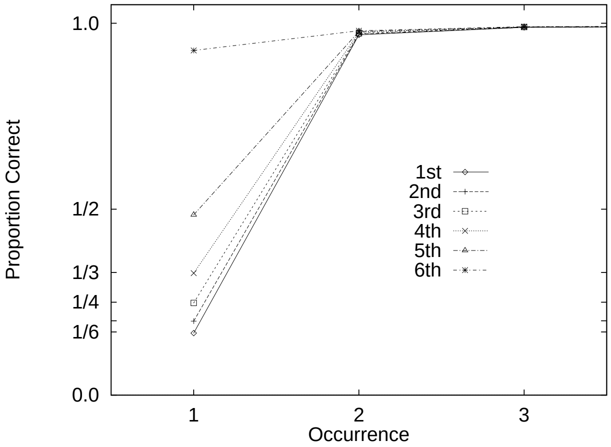

Figure 10. Best fit of the extended model to the results of Experiment 3. The legend indicates the phase of learning.

STM raises performance on the high-frequency diseases in the early trials, and strategic guessing raises performance on the initial exposures to the low-frequency diseases.

Fit to Massed vs. Distributed Learning

The best fit of the extended model to the data from Experiment 2 is shown in Figure 9. Best fitting parameter values were $\lambda = 0.0649$, $\mu = -0.234$, $\sigma = 0.602$, yielding $G^2_{(37)} = 45.09$. The model cannot be rejected statistically, $p > 0.10$.

The model is able to show a large difference between massed and distributed training because of its use of short term memory. The second occurrence of massed pairs is far better than the second occurrence of the distributed pairs because STM remains strong for a few trials.

The model also shows the analogue of the classic spacing effect, yielding somewhat better performance on the first occurrence of the distributed pairs than on the first occurrence of the massed pairs. It is able to show that effect also by dint of STM: The first occurrence of a distributed pair was preceded by the same pair six trials earlier, on average. The first occurrence of a massed pair was preceded by the same pair eleven trials earlier. Hence STM benefits the first occurrence of the distributed pairs more than the first occurrence of the massed pairs. Whereas the STM mechanism can reproduce the spacing effect in this context, it is probably not the correct explanation of classic spacing effects for free recall tasks, in which the test trials come long after STM is presumably acting. Therefore, we place more emphasis on the ability of the model to show an advantage for performance on the second occurrence of the massed pair, and put less emphasis on the ability of the model to show an advantage for the first occurrence of the distributed pair.

Fit to Phased Learning

The best fit of the extended model to the data from Experiment 3 is shown in Figure 10. Best fitting parameter values were $\lambda = 0.792$, $\mu = -0.649$, and $\sigma = 0.0$, yielding $G^2_{(15)} = 30.72$. Qualitatively and quantitatively this is a far better fit than the basic delta-rule model, but the model can be statistically rejected, marginally exceeding the critical χ^2 value at the .01 level. One reason the model has a large fit statistic is because it cannot show performance *worse* than chance for the first trial, as the human data do. Eliminating this single datum reduces the fit statistic by 8.77 to $G^2_{(14)} = 21.95$, for which $p > .05$.

The model shows a dramatic improvement in fit relative to the basic delta-rule model because of its use of strategic guessing. The extended model shows a rise in performance on the first occurrence of the new disease in each phase, just as human learners do.

The best fitting parameters values for phased learning included a very high learning rate and no use of STM. The reason for the large learning rate is that STM, as formalized, is not strong enough for the later-phase diseases. For example, in the fifth phase, the second occurrence of the fifth disease typically does not immediately follow the first occurrence; instead the two occurrences are separated from each other by several trials. The model's STM strength for the fifth disease has decreased significantly by its second occurrence, but people show extremely high accuracy. The model can accommodate the high performance to some extent by using its strategic guessing mechanism, but the extremely good performance can be gained only by setting the learning rate to a large value. Why STM is reduced to zero is less clear. Indeed, if the STM parameter σ is fixed at a value of 0.55 —comparable to the best fitting values for Experiments 1 and 2— and ϕ and μ are allowed to vary, the best fit is not significantly worse: $G^2_{(16)} = 32.37$ for $\lambda = 0.694$ and $\mu = -0.468$ (the increase in G^2 is 1.65, which is not significant for an increase of 1 degree of freedom).

General Discussion

We have shown that the basic delta-rule model cannot fit quantitatively human learning curves for associations that occur with different relative frequencies, despite the recent application of the model to several experiments involving differential base rates. We suggested that at least two mechanisms are involved in human performance that are not reflected in the basic delta-rule model, namely, short-term memory and strategic guessing. Two experiments showed robust effects attributable to STM and strategic guessing, respectively. An extended model formalized the notions of STM and strategic guessing, and provided much better fits to the data.

Inadequacies of the Formalization

As stated earlier, the primary goals of the modeling are to show that the delta-rule model can be greatly improved by adding mechanisms of STM and strategic guessing, and — by virtue of the extended model fitting the human data — to provide further evidence that those additional mechanisms are being used by human learners. We make no claims that the particular formalisms that we employed are necessary. Indeed, the formalisms suffer some inadequacies.

One deficiency is that the model specifies no mechanism, connectionist or otherwise, for interference in STM. Specifying this mechanism might greatly enhance the scope of the model. In particular, Experiment 3, on phased learning, showed that the rigid decrease in STM was inappropriate, and had to be compensated by a large learning rate for long-term associations. It may be the case that the limited capacity of STM is strategically managed, so that STM is used more actively for those associations that are not yet strong in long-term memory, but STM suffers relatively little interference from exposure to items already strong in long-term memory.

The mechanism for strategic guessing suffers at least two problems. First, the guessing strategy is applied, full force, beginning with the very first trial. Humans, however, may instead apply the strategy only after experiencing a few trials and inducing and verifying the underlying assumption of the strategy, namely, that there is a one-to-one correspondence between symptoms and diseases. Non-deterministic or many-to-one mappings might extinguish the strategy. The problem is that the model has no meta-learning mechanism that invents guessing strategies in response to experience, as people do. A second problem is that the model incurs no penalty in capacity or learning for applying the guessing strategy, but there probably is a cost for human learners. The cost of application was reflected in the model by applying the guessing strategy only to the extent that there was uncertainty (entropy) in the initial prediction.

Ramifications

The Rescorla-Wagner model of associative conditioning in animals is formally equivalent to the delta rule for weight changes[5] and that equivalence was an important motivation for the research of Gluck and Bower (1988). The Rescorla-Wagner model has been used to address a variety of findings in animal learning, with mixed success (For recent reviews, see Miller, Barnet, & Grahame, 1995; Pearce, 1994). It is natural to ask, therefore, whether animals show the same effects of STM and strategic guessing as adult humans, or whether the basic delta rule is adequate to describe animal performance in situations comparable to the experiments reported here.

We have assumed that strategic guessing is based on a meta-cognitive strategy, through which the learner uses knowledge of the one-to-one correspondence of symptoms and diseases to infer that unknown symptoms should map to unknown diseases. Other mechanisms might lead to the same effect, however. For example, it might be that there is a bias to choose novel responses for novel stimuli, regardless of any assumptions about one-to-one correspondences. Future research could try to distinguish the hypotheses by independently manipulating novelty and the mapping of symptoms to diseases. Another approach might be to measure the performance of young children, whose meta-knowledge skills are less developed than adults (e.g., Lovett & Flavell, 1990).

Previous researchers have noted that learning may be driven in part by the perceived novelty of the stimuli. Shanks (1992) formalized the ideas of Wagner (1978) so that cues that occurred less often, i.e., cues with greater novelty, developed larger learning rates than frequent cues. That might reflect a real process in human and animal learning, but it does not help account for the results of the experiments reported here. The modification cannot improve performance on the *first exposure* to rare cues, because it affects only the learning rate and not the choice in the absence of knowledge. And contrary to STM, the modification works *against* improvements in performance on repeated consecutive exposures by reducing the learning rate for frequently occurring cues.

Mechanisms of STM and strategic guessing were almost certainly used by participants in previously published experiments. In studies of the inverse base-rate effect, Kruschke (1995) found that human learners performed significantly above chance on their first exposure to a rare symptom, suggesting that strategic guessing was employed. In a study of frequency effects on stimuli that varied on two continuous dimensions, Kruschke and Erickson (1994) found results much like those of Experiment 1 in this article: Early performance on the high-frequency stimuli, and intermediate performance on the low-frequency stimuli, rose much faster than predicted by the delta rule. Future work will test the adequacy of the modeling formalisms to account for those data. A goal for the near future is to combine the principles of STM and strategic guessing described in this article, with the principles of rapid attention shifts and base-rate bias described by Kruschke (1995), along with the principles of

[5] The Rescorla-Wagner model does not use the Luce choice rule to map response strengths to choice probabilities, unlike the delta-rule model used in this article, and the Rescorla-Wagner model was designed to apply to stimuli with multiple cues, unlike the single-cue stimuli used in the experiments reported in this article.

error-driven learning and exemplar-based representation (Kruschke, 1992, 1993b), into a more comprehensive model that addresses thoroughly both the learning and transfer phases of category learning.

References

Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Incorporating prior biases in network models of conceptual rule learning. *Memory & Cognition*, *21*, 413–423.

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchugen zur Experimentellen Psychologie.* Leipzig: Dunker & Humboldt. (Translated by H. A. Ruger and C. E. Byssenine (1913), *Memory: A contribution to experimental psychology*. New York: Dover)

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*, 556–576.

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

Hintzman, D. L. (1976). Repetition and memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 10, pp. 47–91). San Diego: Academic Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Kruschke, J. K. (1993a). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3–36.

Kruschke, J. K. (1993b). Three principles for models of category learning. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), *Categorization by humans and machines: The psychology of learning and motivation* (Vol. 29, pp. 57–90). San Diego: Academic Press.

Kruschke, J. K. (1995). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *00*, 000–000. (in press)

Kruschke, J. K., & Erickson, M. A. (1994). Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model. In *The proceedings of the sixteenth annual conference of the cognitive science society* (pp. 514–519). Hillsdale, NJ: Erlbaum.

Lewandowsky, S. (1995). Base-rate neglect in ALCOVE: A critical reevaluation. *Psychological Review*, *102*, 185–191.

Lovett, S. B., & Flavell, J. H. (1990). Understanding and remembering: Children's knowledge about the differential effects of strategy and task variables on comprehension and memorization. *Child Development*, *61*, 1842–1858.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988). *Journal of Experimental Psychology: General*, *118*, 417–421.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, *117*, 363–386.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*, 352–369.

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 28, pp. 207–250). San Diego: Academic Press.

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

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, *101*, 587–607.

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

Schneider, W. (1988). Micro Experimental Laboratory: An integrated system for IBM PC compatibles. *Behavior Research Methods, Instruments, & Computers*, *20*, 206–217.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology*, *42A*, 209–237.

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

Shanks, D. R. (1991b). A connectionist account of base-rate biases in categorization. *Connection Science*, *3*, 143–162.

Shanks, D. R. (1992). Connectionist accounts of the inverse base-rate effect in categorization. *Connection Science*, *4*, 3–18.

Wagner, A. R. (1978). Expectancies and the priming of STM. In S. H. Hulse, H. Fowler, & W. H. Honig (Eds.), *Cognitive processes in animal behavior* (pp. 177–210). Hillsdale, NJ: Erlbaum.

Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological Review*, *72*, 89–104.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences.* Hillsdale, NJ: Erlbaum.