Title

Mechanisms of long-term repetition priming and skill refinement: A probabilistic pathway model

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 25(25)

ISSN

1069-7977

Authors

Mozer, Michael C.
Colagrosso, Michael D.
Huber, David E.

Publication Date

2003

Peer reviewed

Mechanisms of long-term repetition priming and skill refinement: A probabilistic pathway model

Michael C. Mozer[+*], Michael D. Colagrosso[+*], David E. Huber[#*]
[+] Department of Computer Science
[#] Department of Psychology
[*] Institute of Cognitive Science
University of Colorado, Boulder, CO 80309
{mozer,colagrom,dhuber}@colorado.edu

Abstract

We address an omnipresent and pervasive form of human learning—*skill refinement*, the improvement in performance of a cognitive or motor skill with practice. A simple and well studied example of skill refinement is the psychological phenomenon of *long-term repetition priming*: Participants asked to identify briefly presented words are more accurate if they recently viewed the word. We simulate various phenomena of repetition priming using a probabilistic model that characterizes the time course of information transmission through processing pathways. The model suggests two distinct mechanisms of adaptation with experience, one that updates prior probabilities of pathway outputs, and one that increases the instantaneous probability of information transmission through a pathway. These two mechanisms loosely correspond to bias and sensitivity effects that have been observed in experimental studies of priming. The mechanisms are extremely sensible from a rational perspective, and can also explain phenomena of skill learning, such as the power law of practice. Although other models have been proposed of these phenomena, we argue for the probabilistic pathway model on grounds of parsimony and the elegant computational perspective it offers.

Acquisition of a cognitive or motor skill occurs in several stages. First, an individual must learn the conceptual structures required for the task, including the basic knowledge necessary to perform the task. Then, over a long period of practice, the skill is refined, leading to more fluent, efficient, and robust performance. Skill refinement is an omnipresent and pervasive form of learning. Although skill refinement is sometimes deliberate, e.g., rehearsing a musical piece, it is often implicit, e.g., when typing, driving, reading, playing video games, etc. Understanding skill refinement is fundamentally about discovering the mechanisms by which one trial or performance of the skill leads to improvements on the next.

Long-term repetition priming

Perhaps the most direct and easily studied manifestation of skill refinement in the psychological literature is the phenomenon of *long-term repetition priming*. In the priming paradigm, participants engage in a series of experimental trials, and experience with a stimulus or response on one trial results in more efficient processing on subsequent trials. Efficiency is characterized in terms of faster response times, lower error rates, or both. A typical long-term perceptual priming experiment consists of a *study phase* in which participants are asked to read a list of words one at a time, and a *test phase*, during which they must respond to a series of brief, masked target words. Repetition priming occurs when a word from the study phase influences performance during the test phase. These experiments often vary the *flash duration*, the time between target and mask onsets, and also utilize a variety of response paradigms, including speaking the target identity aloud (*naming*) and forced choice identification between two alternatives (*2AFC*).

Priming is an implicit memory phenomenon: participants are not told the study and test phases are related, and they do not try to recall study words during the test phase as a deliberate strategy for performing the task. Thus, priming is incidental to the test phase of the experiment; it comes about as a result of experience and is thus a form of skill refinement, where the "skill" here is perceptual processing of a letter string.

A key question concerning repetition priming is whether priming is due to increased *bias* or increased *sensitivity*. Although the terminology is borrowed from signal detection theory, the meaning of these terms in the context of priming is somewhat different. Bias means that participants are more likely to report studied items regardless of what word is presented for identification. Sensitivity means that participants become better at perceptual discrimination of the studied items. In a word naming task, improved performance following priming could be due either to increased bias or increased sensitivity. Consequently, experimental paradigms have been designed to unconfound these possibilities. The basic result found in the long-term repetition priming literature is that priming reflects both increased bias and increased sensitivity, although the sensitivity increase is robust only for low-frequency words or novel items.

The goal of this paper is to introduce a model of performance and refinement of simple cognitive skills, such as word reading. The model has two distinct learning mechanisms which contribute to skill improvement with practice. The model explains various data from psychological studies of long-term repetition priming. In this paper, we model two experiments isolating bias and sensitivity components to priming, and show a rough correspondence between our two learning mechanisms and these two effects. We compare our model to existing models in the literature; our model shares aspects of ex-

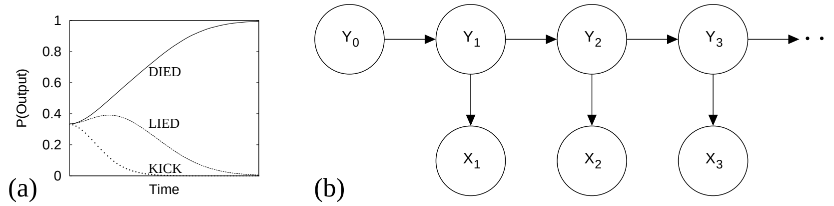

Figure 1: (a) Illustration of a perceptual pathway when the visual word DIED is presented. The three curves show the probabilities of alternative pathway outputs as a function of processing time. In this example, the pathway asymptotes to the correct output with probability 1. (b) An HMM implementation of a pathway

isting models, but has an elegant and concise formulation that makes it preferable on grounds of parsimony.

Modeling long-term repetition priming

The model we present is distilled from a broader theory of cortical information transmission (Colagrosso, 2002). The theory posits that cortical computation is performed by a set of functionally specialized *pathways*. Each pathway performs a primitive cognitive operation, e.g., visual word-form recognition, identification of semantic features of visual objects, computation of spatial relationships, or construction of motor plans. To model the effects of long-term repetition priming, we propose a model with two pathways in cascade: a *perceptual* pathway that maps visual features to word identities, and a *response* pathway that takes the output of the perceptual pathway and maps it to a task-appropriate response. We assume the pathways communicate continuously during processing and that communication is unidirectional.

Pathway as a dynamic belief network

We present a probabilistic model of a pathway, which characterizes the time course of information processing in a single stimulus presentation.

The inputs and outputs of a pathway are represented as probability distributions over distinct alternatives. Formally, the input and output states of a pathway at a particular time t, denoted X_t and Y_t, respectively, are discrete random variables. Each variable can take on one of a finite set of values selected from a multinomial distribution, with set size N_X and N_Y for X_t and Y_t, respectively. We wish to model the temporal dynamics of a pathway, i.e., how X_t and Y_{t-1} combine to determine Y_t. To link this notation to the repetition priming paradigm, consider a perceptual pathway. To model the processing of some word x for a brief duration d, we would set $X_1 = X_2 = \ldots = X_d = x$ (i.e., assigning the random variables a particular value x); to model the masking of the word, X_t for $t > d$ is reset to a uniform distribution over alternatives. Given this input sequence corresponding to a single trial, we can then observe the temporal evolution of the pathway output (Figure 1a).

The relationship among the input and output variables is specified by the graphical model in Figure 1b, known as a *dynamic belief network* (Dean & Kanazawa, 1989; Kanazawa, Koller, & Russell, 1995). Each arrow corresponds to a conditional probability distribution specifying the relationship between two dependent variables. For the reader unfamiliar with graphical models, one should not be concerned with the direction of the arrows. Casting the model as a generative process—where the arrows flow from outputs to inputs in Figure 1b—has certain benefits. Nonetheless, inference can be carried out in either direction: the graphical model allows us to infer the probability distribution over Y_t given $X_1, X_2, \ldots, X_t$, denoted $\mathrm{P}(Y_t|X_1, X_2, \ldots, X_t)$. This computation is performed via iterative Bayesian belief revision. Figure 1b is simply a hidden Markov model (HMM), used in a novel way. In typical usage, an HMM is presented with a sequence of distinct inputs, whereas we maintain the same input for many successive time steps. Further, in typical usage, an HMM transitions through a sequence of distinct hidden states, whereas we attempt to converge with increasing confidence on a single state.

In Figure 1b, the set of arrows from X_t to Y_t corresponds to $\mathrm{P}(X_t|Y_t)$, the *instantaneous transmission probability* between X_t and Y_t. The set of arrows from Y_{t-1} to Y_t corresponds to $\mathrm{P}(Y_t|Y_{t-1})$, and can be thought of as a *short-term memory* in the pathway output. In dynamic belief networks, it is typical to assume temporal invariance of the conditional distributions, i.e., $\mathrm{P}(X_t|Y_t) = \mathrm{P}(X|Y)$ and $\mathrm{P}(Y_t|Y_{t-1}) = \mathrm{P}(Y|Y_{prev})$ for all t. This assumption is equivalent to stating that the parameters of these distributions are *homogeneous*—the relationship between pathway inputs and outputs does not change on the brief time scale of information processing modeled. The two distributions, $\mathrm{P}(X|Y)$ and $\mathrm{P}(Y|Y_{prev})$, embody the knowledge in a pathway. In the following two sections, we discuss these forms of knowledge, which constitute the central claims of the model.

Instantaneous transmission probabilities The instantaneous transmission probability between some $X = i$ (the random variable X taking value i) and some $Y = j$ is formulated as $\mathrm{P}(X = i|Y = j) \sim 1 + \alpha_{ij}$ where α_{ij} denotes the *association frequency*, and is assumed to be related to the number of previous experiences with the association between $X = i$ and $Y = j$. (The probability must be normalized; hence the definition is formulated in terms of a proportionality instead of an equality. The constant 1 prevents renormalization by zero.) Increasing α_{ij} increases the instantaneous probability of information, and thus increases the rate at which information about X is communicated to Y.

Although the input representation is localist, in that there is one value of X for each possible input, one can achieve the similarity structure inherent in a distributed representation using explicit terms, γ_{ik}, that specify the similarity between input states i and k:

$$\mathrm{P}(X = i | Y = j) \sim 1 + \sum_k \gamma_{ik}\alpha_{kj}$$

Short-term memory We assume that the transition probability matrix from Y_{prev} to Y acts as a memory with diffusion. That is, with probability β, Y is reset to its initial state and with probability $(1 - \beta)$, Y remains in the same state as Y_{prev}:

$$\mathrm{P}(Y = i | Y_{prev} = j) = (1 - \beta)\delta_{ij} + \beta \mathrm{P}(Y = i)$$

where β is the diffusion constant, $\mathrm{P}(Y)$ is the *prior* distribution (the output of the pathway in the absence of any input), and δ_{ij} is the Kroniker delta ($\delta_{ij} = 1$ if $i = j$ or 0 otherwise). If $\beta = 0$, the transition matrix acts as a perfect memory.

Processing dynamics The distribution over Y_t given the input sequence, $\mathbf{X}_t \equiv \{X_1, X_2, \ldots, X_t\}$, can be derived from Bayes' theorem, based on the information transmission probabilities, $\mathrm{P}(X|Y)$, the pathway output transition probabilities, $\mathrm{P}(Y|Y_{prev})$, and the prior distribution $\mathrm{P}(Y = k) \equiv \mathrm{P}(Y_0 = k|\mathbf{X}_0)$:

$$\mathrm{P}(Y_t = k|\mathbf{X}_t) \sim \left[\sum_{j=1}^{N_X} \mathrm{P}(X_t = j)\mathrm{P}(X_t = j|Y_t = k)\right] \left[\sum_{i=1}^{N_Y} \mathrm{P}(Y_{t-1} = i|\mathbf{X}_t)\mathrm{P}(Y_t = k|Y_{t-1} = i)\right]$$

To model two pathways in cascade, such as the perceptual and response pathways, the output of the perceptual pathway is provided as input to the response pathway. Although the two pathways could be coupled into a single graphical model, inference in this model is intractable. Consequently, we approximate inference by assuming that at each time step the perceptual pathway output is copied to the response pathway input. This decoupling corresponds to the assumption of limited communication between pathways.

Learning mechanisms

Our simulations assume that the model has already acquired the basic knowledge necessary to perform a task. Thus, the α_{ij} are initialized such that the model produces the correct association asymptotically. Because the α_{ij} are presumed to reflect the frequency with which an association has been exercised in the past, we initialize high-frequency associations to have larger values.

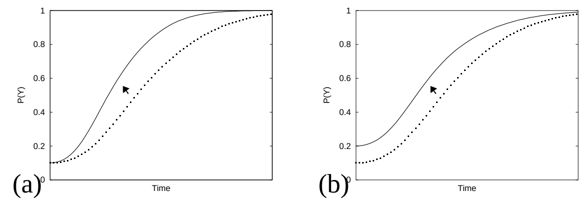

Figure 2: Change in the time course of activation of a pathway resulting from (a) an adjustment to the instantaneous transmission probabilities and (b) an adjustment to the priors

How does experience affect parameters of the model? Based on the parameter definitions, two sets of parameters might logically be adapted: the association frequencies and the priors. By definition, α_{ij} reflects the frequency that an association has been experienced. Consequently, it should increase with each experience. Also by definition, the output priors, $\mathrm{P}(Y)$ should reflect statistics of the environment. In a normative model, these statistics should be updated over trials to accommodate nonstationary and unknown environments.

For the association frequencies, we chose the simple rule in which α_{ij} is incremented by the constant $\Delta\alpha$ following a trial in which input i leads to activation of output j. To form an analogous rule for the prior, we define the prior $\mathrm{P}(Y = j)$ in terms of secondary parameters, $\mathrm{P}(Y = j) \sim 1 + \rho_j$, and update ρ_j by the constant $\Delta\rho$.

These two adaptation mechanisms cause the transmission of signals to become more efficient with experience, i.e., pathway accuracy increases given a fixed amount of processing time, or pathway response time decreases to achieve a desired level of accuracy. Efficiency is reflected in a leftward shift of the curve relating processing time to output probability. The association-frequency update results in a more rapid integration of the output probability (Figure 2a); the prior update raises the initial probability of the response (Figure 2b).

Simulations

We explain data from two key studies of long-term repetition priming. The purpose of presenting these simulations is establish the plausibility of the model. Other models have been developed to explain the same phenomena, although our model can explain a broader range of data (not presented in this paper). We we will conclude by discussing reasons to prefer our model.

Our simulations utilized two pathways which were identical, except for the diffusion rates in the perceptual and response pathways, $\beta^p = 0.05$ and $\beta^r = 0.01$, respectively. The pathways were designed to produce 1-1 mappings, with $N_X = N_Y = 20$ input and output states for each pathway. We assume a similarity structure in which each input state is reciprocally similar to two others with a uniform similarity coefficient $\gamma = 0.8$. Rather than assuming independent α_{ij} parameters for each association, we used two values, one for high-frequency

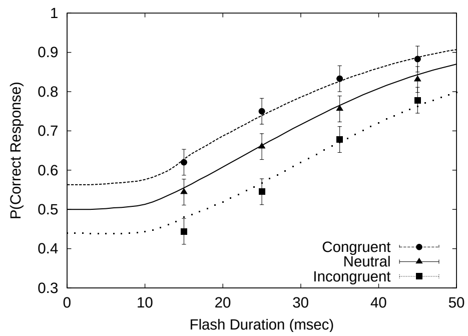

Figure 3: Accuracy of response for congruent, incongruent, and neutral conditions of the Ratcliff & McKoon (1997, Experiment 3) study of bias effects in priming. The points are results from human subjects, and the curves are produced by our model.

words, $\alpha_{\text{high}} = 7.5$, and one for low-frequency words, $\alpha_{\text{low}} = 1.25$, with $\Delta\alpha = .625$. Finally, the prior update rule has one free parameter, $\Delta\rho = 3.3$. In total, the model had seven independent parameters, although the model's behavior was insensitive to the exact parameter values. One additional constraint was that we chose parameters such that one simulation time step corresponds to one millisecond in the experimental studies.

Simulation 1: bias effect

One explanation for the facilitatory effect of repetition priming is that study of the prime introduces a response bias that increases the probability of reporting the prime in the future. Ratcliff and McKoon (1997, Experiment 3) explored the bias account of priming in a 2AFC paradigm. During the test phase, masked target words were briefly presented, followed by a two alternative forced choice between the target and a distractor alternative. The target and distractor were orthographically similar, making the discrimination more difficult.

Three experimental conditions were contrasted: In the *congruent* condition, the target was presented during the study phase. In an *incongruent* condition, the distractor was presented during the study phase. In a neutral condition, neither was previously studied. For example, if DIED was studied, then target DIED with distractor LIED would be a congruent trial; target LIED with distractor DIED would be an incongruent trial; and target KICK with distractor SICK would be a neutral trial. The experiment also manipulated the *flash duration*, the asynchrony between target and mask onset.

Human performance in the experiment is indicated by the points and error bars in Figure 3 for flash durations of 15, 25, 35, and 45 msec. Across flash durations, the accuracy benefit on congruent trials relative to neutral trials is matched by an accuracy cost on incongruent trials, diagnostic of a bias effect.

Our model produces an excellent fit to the data, as shown by the curves in Figure 3. The stimulus presentation durations are modeled by setting the perceptual pathway input distribution such that $\mathrm{P}(X_t = i) = 1$ for stimulus i over the flash duration, and following that time, resetting the pathway input to the state of no information, the uniform distribution. Figure 4a illustrates the operation of a perceptual pathway for the stimulus DIED with a flash duration of 25 msec. Following the removal of the stimulus, the perceptual pathway decays back to its prior distribution at a rate proportional to β^p. In this example, DIED has been previously studied, as indicated by the fact that DIED has a higher probability at $t = 0$ than any other word. The response pathway, shown in Figure 4b, accumulates evidence from the perceptual pathway, reaching an asymptote as the perceptual pathway decays. To produce a 2AFC response, we adopt the normative assumption that the 2AFC response is computed from the response pathway output, conditional on the output being one of the two response alternatives.

Although two mechanisms of adaptation are built into the model, the prior update rule is almost entirely responsible for the differences in performance among conditions. Setting the association-frequency adjustment, $\Delta\alpha$, to zero has little impact on the simulation results. Thus, the prior update rule roughly corresponds to the notion of bias. Indeed, one can see this bias manifested at $t = 0$ in Figure 4a. However, with increased flash duration, the probability of correct identification of the target approaches asymptote, the differences among conditions diminish, and the bias disappears. A simple rule that adjusted response probabilities independent of flash duration could not account for the data; the temporal dynamics of the model are essential to explaining the phenomenon.

Simulation 2: sensitivity effect

As a complement to the bias effect, the sensitivity effect is an improvement in perceptual discrimination of an item as a result of previous study. To determine if a sensitivity effect contributes to priming, several studies (Bowers, 1999; McKoon and Ratcliff, 2001, Experiment 2; Wagenmakers, Zeelenberg, and Raaijmakers, 2000) explored a 2AFC task in which a comparison is performed between a condition in which both response alternatives are primed and a condition in which neither response alternative is primed (*both* and *neither primed*, respectively). Any difference between these conditions could not be attributed to a bias effect, because—by the simple notion of bias—the bias effect should cancel when both alternatives are primed. A reliable benefit in the both-primed condition relative to the neither-primed condition is therefore diagnostic of a sensitivity effect. We model the study of (McKoon & Ratcliff, 2001), which included a word frequency manipulation.

The human data shows a reliable benefit of study for low-frequency words (e.g., WOMB, TWIG), but not for high-frequency words (BEEN, THAN), as shown in Figure 5a. (The difference between both and neither conditions for high frequency words is nonsignificant.) Thus, priming can improve the discriminability of low-

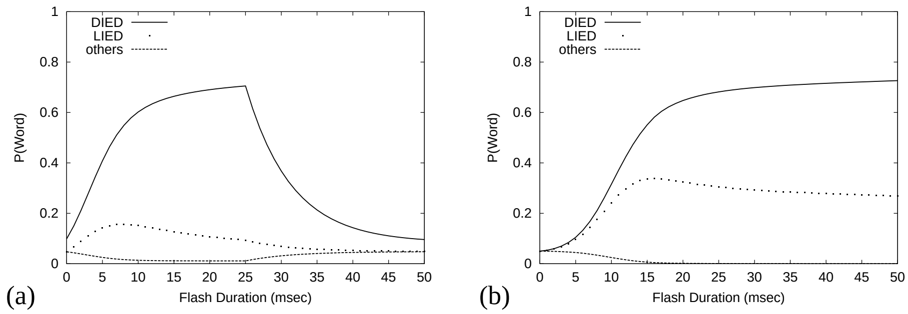

Figure 4: Output of the perceptual and response pathways (left and right panels) for a 25 msec presentation of the target DIED on a congruent trial

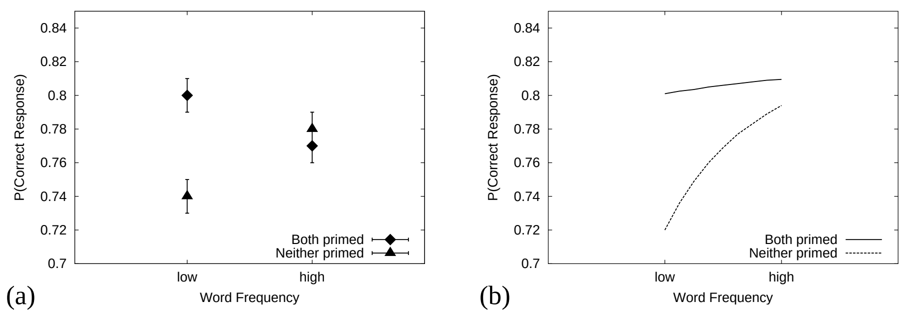

Figure 5: 2AFC accuracy for both- and neither-primed conditions, for low- and high-frequency words. (a) human data from McKoon and Ratcliff (2001); (b) simulation results from our model.

frequency words.

Our model produces the same qualitative pattern as the experimental data. The benefit of study diminishes with word frequency, as reflected in the convergence of the two curves in Figure 5b. This result is due to the adjustment of association frequencies: if the adjustment of priors is turned off by setting $\Delta\rho = 0$, the qualitative pattern of results in Figure 5b is unaffected.

Because of the assumption that α is proportional to frequency, each experience with an item must result in a fixed increment to α. However, the fixed increment has a greater effect on performance for small α than large α, due to the normalization of the conditional transmission probabilities: for the correct association, $\mathrm{P}(X|Y) = \alpha/(\alpha + N_X)$. The derivative of this expression, $\partial\mathrm{P}(X|Y)/\partial\alpha = N_X/(\alpha + N_X)^2$, specifies the boost in transmission probability with a fixed increment in α. The derivative drops quadratically with α, but the effect on performance is even greater because this transmission probability influences the model's output on each of hundreds of time steps. A simple simulation shows the frequency effect more clearly. Figure 6 simulates the accuracy for a fixed stimulus presentation duration as a function of α. For equal increments in α, the accuracy gain is greater for a low frequency word than for a high frequency word. Figure 6b is a log-log plot of accuracy versus α. The straight line indicates a power law of practice emerging from the model. (A plot of response time versus α yields the same result.)

Discussion

We described a probabilistic model that offers a compact, formal language for characterizing the time course of information transmission, and the changes in information transmission due to long-term repetition priming. The model explains key phenomena in the long-term repetition priming literature, including: the bias and sensitivity effects, the dependence of the sensitivity effect on word frequency, and the time course of priming within a trial. We (Colagrosso, 2002; Mozer, Colagrosso, & Huber, 2002) have used this model to address other priming phenomena, including: the effects of target-distractor similarity, the decay of bias effects over time, alternative response paradigms including naming and matching, and response priming effects. The elegance of the model stems in part from the Bayesian framework, which dictates the mechanisms of inference within a pathway, and in part from parameters that correspond directly to quantities of psychological interest, such as interitem similarity (γ) and degree of experience (α).

Other models have been proposed to explain the data addressed in in this paper, most notably REM (Schooler, Shiffrin, & Raaijmakers, 2001) and the counter model (McKoon & Ratcliff, 2001; Ratcliff & McKoon, 1997). Our model has some similarities with these models: we share the assumption with REM that perceptual and

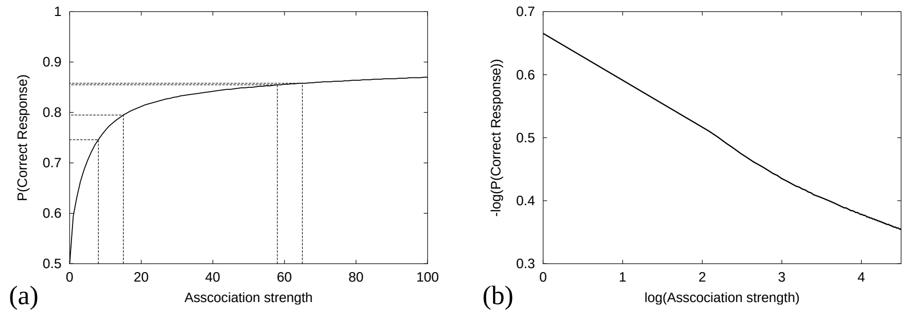

Figure 6: (a) The accuracy of the model as a function of association frequency, α, for a fixed flash duration. (b) A log-log plot of accuracy versus association frequency.

memory systems adapt to achieve optimal performance over evolution and development; and we share the assumption with the counter model of gradual accumulation of decision-making evidence over time. However, the existing models have some serious weaknesses. REM makes the unparsimonious assumption that an item's lexical trace is augmented with context, which allows the model to behave as if it is taking into account prior probabilities, whereas we model the priors directly. Further, although REM is based on a probabilistic framework, it gets much less leverage from the framework than does our model. The counter model operates in a currency of counts, and the rules for accruing counts are somewhat arbitrary, e.g., the stealing of counts by a studied item from visually similar neighbors. Neither model offers a natural explanation for increased sensitivity to low-frequency words. And most importantly, neither model has intrinsic temporal dynamics that lead to strong predictions concerning performance as a function of stimulus exposure duration.

Our model has two virtues. First, despite its parsimony, it offers a broad conceptual framework, not restricted to a particular experimental paradigm or task. Second, the two distinct mechanisms that explain bias and sensitivity effects were introduced not simply to explain the data, but are motivated on logical grounds, in contrast to the existing models. The mechanisms—one that adjusts the pathway output prior probabilities and the other that adjusts transmission probabilities within a pathway—are extremely sensible mechanisms for an adaptive system. The priors can be viewed as a simple model of the environment, and updating this model is appropriate if encountering an object in one's environment implies that one is more likely to encounter the object in the future. The transmission probabilities can be viewed as a limited-capacity resource, and allocating this resource to recently performed cognitive operations is judicious assuming that they are likely to be required again. These primitive mechanisms subserve not only long-term priming, but also offer insight into the more general phenomenon of skill refinement. One example of this claim is the power law of practice the model exhibits, ubiquitous in human performance. Another example is the prediction of the model that associative strengthening due to priming should be longlasting and association specific. These properties appear to be robust characteristics of skill learning (Healy & Bourne, 1995).

Acknowledgments

This research was supported by Grant 97–18 from the McDonnell-Pew Program in Cognitive Neuroscience, NSF award IBN–9873492, and NIH/IFOPAL R01 MH61549–01A1.

References

Bowers, J. S. (1999). Priming is not all bias: Commentary on Ratcliff and McKoon (1997). *Psychological Review*, *106*, 582–596.

Colagrosso, M. D. (2002). *A rational theory of skilled performance and practice: Modeling long-term repetition priming.* Unpublished doctoral dissertation, University of Colorado.

Dean, T., & Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Artificial Intelligence*, *93*, 1–27.

Healy, A. F., & Bourne, L. E. (1995). *Learning and memory of knowledge and skills: Durability and specificty.* Thousand Oaks, CA: Sage.

Kanazawa, K., Koller, D., & Russell, S. J. (1995). Stochastic simulation algorithms for dynamic probabilistic networks. In *Proceedings of the 11th Annual Conference on Uncertainty in AI (UAI)* (pp. 346–351). Montreal, Canada.

McKoon, G., & Ratcliff, R. (2001). Counter model for word identification: Reply to Bowers (1999). *Psychological Review*, *108*, 674–681.

Mozer, M. C., Colagrosso, M. D., & Huber, D. E. (2002). A rational analysis of cognitive control in a speeded discrimination task. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14.* Cambridge, MA: MIT Press.

Ratcliff, R., & McKoon, G. (1997). A counter model for implicit priming in perceptual word identification. *Psychological Review*, *104*, 319–343.

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review*, *108*, 257–272.

Wagenmakers, E.-J. M., Zeelenberg, R., & Raaijmakers, J. G. W. (2000). Testing the counter model for perceptual identification: Effects of repetition priming and word frequency. *Psychonomic Bulletin and Review*, *7*, 662–667.