Optimizing Knowledge Component Learning Using a Dynamic Structural Model of Practice

Philip I Pavlik Jr. (ppavlik@cs.cmu.edu), Human Computer Interaction Institute
Nora Presson (presson@cmu.edu), Psychology Department
Kenneth Koedinger (koedinger@cmu.edu), Human Computer Interaction Institute
Pittsburgh Science of Learning Center
Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213 USA

Abstract

This paper presents a generalized scheme for modeling learning in simple and more complex tasks, and shows how such a model can be applied to optimizing conditions of practice to maximize some desired performance. To enable this optimal allocation of lesson time, this paper describes how to quantify the preferences of students using utility functions that can be maximized. This conventional game theoretic approach is enabled by specifying a mathematical model that allows us to compute expected utility of various student choices to choose the choice with maximal expected utility. This method is applied to several educational decisions that can benefit from optimization.
Keywords: Memory; Economics; Practice; Computer-Aided Instruction.

Introduction

This paper describes a method for applying economic principles in order to allocate the scarce resource of learning time toward satisfying the unlimited need for education. To do this, we describe a model that decomposes learning into individual knowledge components (KCs) that possess some degree of independence from other skills (a "knowledge component" is any proficiency that can be learned). By assuming this independence, the model accounts for the unique effects of practice on specific KCs, with the goal of optimizing the benefit of practice.

We do not argue that the model is a precise representation of all the processes involved in learning, but rather that it provides a heuristic tool to track observed strengths of KCs as a general function of practice, so that improvement over time and across KCs can be optimized. The model we will present, like similar models, is effective in capturing practice effects (Cen, Koedinger, & Junker, 2006). Further, it is interesting to note that the dynamic practice model presented here (based on the ACT-R computational model of declarative memory, Anderson & Lebiere, 1998) might be substituted with another model of cognition with only minimal modification to the approach.

Although the model is a simplification of learning processes in most cases, this simplicity provides an important advantage in application. It allows closed form predictions of which learning events (LEs) might be assigned at what times to maximize learning (a "learning event" is any discrete interval over which a learned proficiency increases). Ultimately, it is explaining this collection of closed form predictions and recommendations that is the goal of this paper.

To explain these concepts this paper has three parts. The first section on the dynamic practice model is largely a review of the ACT-R model of declarative memory. This section serves to orient the reader on the output functions (probability and latency of recall) that will be used later. The second section on structural models details how compound events can be modeled using the dynamic practice model. Compound events are important to consider when responses are not independent and are especially relevant for certain kinds of optimization situations (i.e. part-task to whole-task transfer of performance). The final section shows several ways the model built in the first half of the paper can be applied to optimizing knowledge component learning.

Dynamic Practice Model

To understand the quantitative model that will be used to predict and optimize learning, we will begin with the equations that predict probability of correct performance and latency of correct performance as a function of the activation strength of a KC.

Probability Correct. The first dependent measure of KC performance is probability of correct response. Equation 1 shows the standard Boltzmann equation (similar to the Rasch model used in item response theory), a logistic function that characterizes the threshold of correct performance (the level of activation at which performance is correct greater than 50% of the time) and distributional noise as τ and s respectively. Equation 1 describes a model of the probability of giving a correct response (p) for a given KC activation strength value (m) and the parameters described above.

$$p_m = \frac{1}{1 + e^{\frac{\tau - m}{s}}} \qquad \text{Equation 1}$$

Latency. A second dependent measure used to track KC performance is latency (labeled q in our model). Various sources suggest modeling latency with a Weibull distribution (Anderson & Lebiere, 1998; Logan, 1995). Such a Weibull distribution can be produced by using Equation 2 to represent latency as a function of F (which scales latency magnitude), m (memory strength) and a fixed cost (which is determined from data and captures the minimum time necessary for perceptual and motor costs of responding). Logistic noise on m determines the shape of the aggregate Weibull function for a population of response latencies.

$$q_m = Fe^{-m} + fixedtimecost \qquad \text{Equation 2}$$

Knowledge Component Strength Function

Given these two output functions, which correspond to two important ways of measuring KC performance, we can now elaborate how current m is computed as a function of the history of a student's practice of a KC practice item across n prior LEs. Equation 3 shows this KC strength function. The history term, the final portion of Equation 3, is essentially described by three values, t, d and b, for each LE. The values for t represent the times since each past LE (the ages of each LE effect). The d values are the power law decay values for each LE. The b values scales the effect of each LE depending on the amount of learning for the LE (i.e. longer duration LEs and successful test LEs result in higher bs). To model history, the bt^{-d} quantity is summed for each of the n learning events (LEs). The logarithm serves to scale the quantity from $-\infty$ to ∞. This power law decay formulation was first explored by Anderson and Schooler (1991), who showed that it results in patterns of forgetting that match the relative need for performance in the environment. The β parameters, the first portion of Equation 3, capture naturally occurring error when the model is fit to data from multiple students or multiple KCs. β_s is the parameter that captures consistent error across KCs as a function of student. β_i captures consistent error across students as a function of KC (i stands for item). Finally β_{si} captures the residual error for a specific KC and a specific student over multiple LEs.

$$m_n = \beta_s + \beta_i + \beta_{si} + \ln\left(\sum_{k=1}^{n} b_k t_k^{-d_k}\right) \qquad \text{Equation 3}$$

$$b_{study} = g(1 - e^{-v \cdot studyduration}) \qquad \text{Equation 4}$$

Equation 4 shows how b can be computed as a function of the duration of a study LE (where v and g represent a growth constant and the maximum possible encoding respectively). This captures the notion that continuous time spent on a single KC has a diminishing effect on learning (Metcalfe & Kornell, 2003). Recent work by Pavlik (in press) has shown how this b scalar can be used to capture the learning difference between active correct responding and passive study. In such work, $b_{successfulretrieval}$ is typically set at a constant, whereas b_{study} varies as described in equation 4. This supposes two canonical forms of the LE: the "study LE," which comes from unassessed study over some fixed period of time of a stimulus representing a KC, and the "test LE," which comes from a variable-duration assessment of learning (test LEs are often followed by study opportunity and then are called "drill LEs"). Test LEs are interesting not only because they tend to lead to more learning than passive study (for correct responses), but also because they provide information about the current state of learning that can be used to implement knowledge tracing. Such knowledge tracing algorithms have changed form over different applications of the model. In the original version (Pavlik Jr., 2005), the distribution of residual β_{si} variance is used as the initial Bayesian prior for item strength and numerical integration is used to adjust this value after each practice by integrating the logistic distribution for correctness given the response of the student. In the more recent version, we have found that a more computationally inexpensive model that allows the simpler $b_{successfulretrieval}$ parameter to capture the β_{si} variance works well in practice (Pavlik Jr. et al., 2007). Further, the latest version also uses a $b_{latency}$ parameter multiplied by each $b_{successfulretrieval}$ parameter for each successful test. This $b_{latency}$ parameter is a natural log transform (with a scalar parameter) of the difference between q_m (the predicted latency) and the latency data from the student. This creates a knowledge tracing model that assumes that faster responding means more learning has occurred.

Equation 5 shows a more recent modification of the ACT-R equations to capture the spacing effect, the spacing-by-practice interaction, and the spacing-by-retention interval interaction (Pavlik Jr. & Anderson, 2005). This change says that the forgetting rate from any LE depends on the level of activation at the time of the LE. As modeled in Equation 5, when spacing between trials gets wider, activation decreases between presentations; decay is therefore less for each new presentation, and long-term probability of correct performance does not decrease as much.

In Equation 5, the decay rate d_k is calculated for the kth presentation of a KC item as a function of the activation m_{k-1} at the time the presentation occurred (e.g., the decay rate for the 7th LE (t_7) depends on the activation at the time of the 7th LE, which is a function of the time from last exposure of the prior 6 LEs and their decay rates. It is important to note that since t_ks are ages (or differences between the current time and the time of the past trial), activation and decay depend on the current time as well as the number of LEs).

$$d_{m_{k-1}} = ce^{m_{k-1}} + a \qquad \text{Equation 5}$$

Anderson, Fincham, and Douglass (1997) found that Equation 3 could account for practice and forgetting during an experiment, but it could not fit retention data over long intervals. Because of this, they concluded that between sessions, the presence of intervening events erodes KCs more slowly than during an experimental session. This slower forgetting was modeled by scaling time as if it were slower outside the experiment. Forgetting is therefore dependent on the "psychological time" between presentations, rather than the true intersession interval. This factor is implemented by multiplying the portion of time that occurs between sessions by h (a scalar parameter for time) when calculating recall. This is done by subtracting h*total intersession time* from each age (t_k) in Equation 11 (Pavlik Jr., 2005; Pavlik Jr. & Anderson, 2005). Because of this mechanism, time in the model is essentially a measure of destructive interfering events. The decay rate, therefore, is a measure of "fragility" of memories to the corrosive effect of these other events.

This model has the flexibility to capture many varieties of learning and practice effects. To further understand this flexibility, consider the issue of more implicit production rule (procedural) learning in contrast to explicit factual (declarative) learning. This distinction is supported by research from widely distinct theoretical perspectives such as ACT-R and connectionism and is supported by dissociable neural mechanisms (McClelland, McNaughton, & O'Reilly, 1995). We might wonder whether the equations just presented are adequate to capture both knowledge (and KC) types. Specifically, that work implies that declarative learning is both faster (reflected by a larger *b* parameter) and more easily forgotten (reflected by a larger *d* parameter) than procedural learning, and our model can clearly characterize these differences.

Structural Model

The structural level model assumes that few domains are made up of entirely independent KCs, as seems to be implied in the model we just presented. The word "structural" refers to the fact that, because of this lack of independence, the modeler must be concerned with the structure that links the multiple KCs and their association. In many domains, predictions of the probability of correct response and latency are derived from the strength of more than one underlying KC. For example, in studies of Chinese vocabulary learning, stimuli can be presented in one of four modes (Hanzi character, pinyin text, sound file, and English text). This results in 6 possible test LE types, two of which are English→pinyin and Hanzi→pinyin drill LEs [(stimulus)→(response)]. In both of these cases, drill success depends not only on the strength of the link between the stimulus and response, but also on the ability to recall and produce the pinyin response. Because of this, performance for these pairs cannot be independent.

Similarly, in work with a French gender identification task, words fall into gender categories based on spelling and semantic cues. For instance, words that end in –age are most often masculine in French, as in *le fromage*. Although each of these words might yield a correct response independent of the general rule (through recall), it is also obvious that all rule exemplars share a KC that can be used to respond to any items in a cue category (and in fact, it is this generalized responding, rather than exemplar-based recall, that we want to optimize).

To deal with the fact that multiple KCs are required for these single skills, we will propose two basic structural models that account for this, each of which fits some possible learning tasks: the conjunctive structure and the disjunctive structure.

Conjunctive Model

In a conjunctive model, all component KCs must be active to produce a correct response. For instance, in the Chinese vocabulary work, probability of correct performance for each trial is captured by the probability of correct recall for both the response and the link between the stimulus and the response. Given this model, probability correct depends on both the strength of the link and the strength of the response in a conjunctive function: p(link) * p(response), such that both elements are necessary for a correct response. The more general form for the conjunction of 2 KCs is shown in Equation 6. Latency, on the other hand, is handled as the sum of the perceptual motor costs, the cost for recall of the link KC, and the cost for recall of the response KC. Not only does this structural model handle the pinyin response example above, but it also captures data showing that responding with a word in the native language should be easier than the recently learned foreign equivalent (e.g. Schneider, Healy, & Bourne, 2002).

$$\mathrm{p}(KC_1 and KC_2) = \mathrm{p}(KC_1)\,\mathrm{p}(KC_2) \qquad \text{Equation 6}$$

Disjunctive Model

The disjunctive model, in contrast, assumes that a trial can yield a correct response due to performance of any one of the two or more independent KCs. Often disjunctive models apply in a generalization situation where the domain contains specific KCs that apply for individual stimuli and general KCs that each apply to a group of stimuli, as in the French gender case. In this example, we can imagine that general group KCs control performance for "clusters," the members of which can also be learned by rote. Given the example of a general (rule-based) and specific (rote) component controlling each performance, probability of correct skill performance depends on the strength of both general and specific components in a disjunctive function, p(general) + p(specific) * (1-p(general)), such that (for example) a student could classify a novel word on the sole basis of the general KC. The general form of this model is shown in Equation 7.

$$\mathrm{p}(KC_1 or KC_2) = \mathrm{p}(KC_1) + \mathrm{p}(KC_2)(1 - \mathrm{p}(KC_1))$$

Equation 7

Optimizing Learning

The following procedures describe how one can use the model to compute optimal practice schedules. Usually, we assume that what is being optimized is gain in some long-term measure of learning for a KC or multiple KCs. Although using long-term probability correct as a dependent measure works when we focus on optimizing some global aggregate task (like the optimal total number of practices for an item), we need a different utility function for more dynamic local scheduling (such as picking an item to practice next), in order to formalize preferences for the learning gains from different LE schedules.

Utility Optimizations

We propose to use Equation 8 as the utility function for a LE (where *b* controls the weight of the LE, *t* is the desired retention interval of the LE, and decay (*d*) is a function of the activation (*m*) at the time of practice). Most importantly, Equation 8 does not have the all-or-none property of probability correct (because probability correct is a sigmoid function, it usually approaches 0 or 1). If we tried to use long-term probability correct as our measure of local utility, it would value practice most heavily when it comes near the transition from mostly incorrect performance to mostly correct performance across a sequence of test LEs (those LEs that fall on the intermediate part of the curve). This bias distorts the fact that we are ultimately more concerned with the minimum number of practice trials required to reach a certain long-term retention, not scheduling each practice trial so that it increases percent correct maximally. These goals are actually quite different since long-term percent correct gain from the next practice depends on nearness to long-term floor or ceiling performance, while utility gain is not affected by these bounds. Thus, our utility function maximizes the overall goal by valuing LEs independently of the order they occurred, considering only their unique contributions (a function of strength of encoding, recency, and the decay rate) to the long-term KC strength.

$$u = bt^{-d_m} \qquad \text{Equation 8}$$

We will use Equation 8 as a cardinal utility function: e.g., a .2 increase in strength is half as good as a .4 increase in strength. One reason why this assumption is reasonable is because LEs contribute to KC strength in small increments and these increments are interchangeable, as illustrated in Equation 3. Using a cardinal utility function allows us to directly compare different possible spacings and KC presentation orders, to determine when learning is maximal, given learning history. Further, we assume that this utility equation satisfies the von Neumann and Morgenstern game theoretic axioms of completeness, transitivity, continuity and independence required for comparing expected utility lotteries (Von Neumann & Morgenstern, 1944).

Practice Spacing Optimization (PSO). For each KC and each student, it is useful to decide when it would optimal to repeat a drill LE of that KC. Therefore, we are trying to schedule the LEs under conditions of allocative efficiency. In economics, allocative efficiency is a condition where costs (time spent learning) are allocated in a way that maximizes gains (increases in utility). Taking this parallel to learning theory, we search for the retention interval (for each KC) at which the expected rate of learning utility gain is maximal given a new LE. This is expressed in Equation 9, which calculates the maximum utility gain for a KC as a function of *m* (activation of that KC) and *t* (the target retention interval needed to compute *g* in Equation 8). All the other values are fixed parameters (b_s = success LE weight from Equation 8 if the test LE is successful, b_f = failure LE weight from Equation 8 for the study LE given as review, $-d$ computed from the current *m* (needed with *t*, b_f and b_s to compute *u* values), p_m and q_m estimated for the test LE from Equations 1 and 2, and failure costs estimated from prior data). Because *t* and *m* are the only values that vary in finding the optimum spacing, we can solve for the optimal level of the one given the other. For example, if we know the desired retention interval, we can solve for the max of Equation 9 to solve for the optimal level of activation at

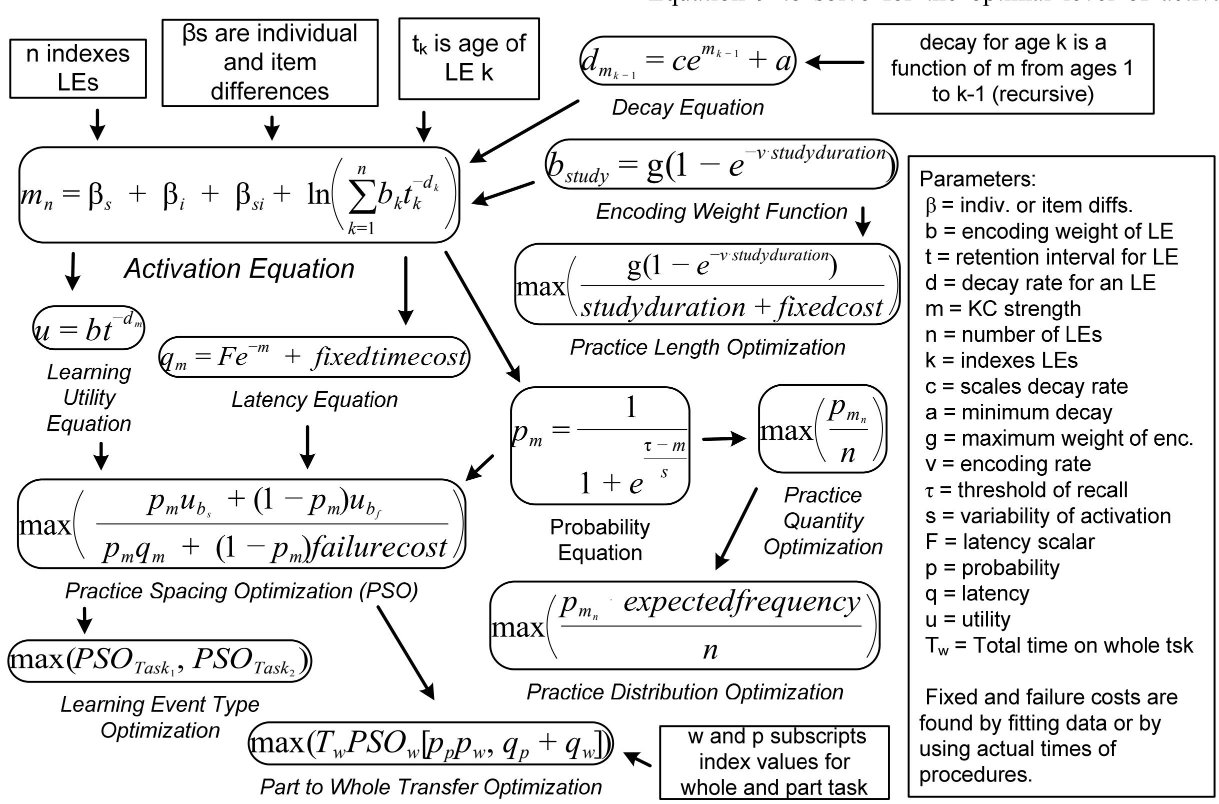

Figure 1. Organizing diagram of the mathematical relationships in this paper.

which practice should occur.

In practice, Equation 9 tends to suggest (for a drill procedure) that when failure costs for errors and error feedback are high, or success gains from correct responding are much greater than failure gains from feedback study, long-term gains in utility per second of practice will be highest when repetitions are scheduled so that test LE performance is maintained at a high probability. However, because the decay parameter can be large for an LE after a short spacing, some spacing is always preferred.

$$\max\left(\frac{p_m u_{b_s} + (1-p_m)u_{b_f}}{p_m q_m + (1-p_m) failurecost}\right)$$

Equation 9

Learning Event Type Optimization. The above discussion assumes a single task (drill) which can be selected for each item. However, we can also propose other types of LEs and then compare them with the drill trial. For example, we could decide whether it was better to give a study LE alone or to give a drill LE(a test LE followed by a study LE when the test fails). To do this, Equation 10 shows how we can compare the learning rates for each trial type to determine the optimal next trial type for the student. This principle can be extended to compare any two tasks (e.g., tutored problem solving vs. untutored problem solving). This is typically used in combination with dynamic PSO calculations (when the *PSO*s in Equation 10 are computed as a function of the current time) to pick the optimal time for the optimal task.

$$\max(PSO_{Task_1}, PSO_{Task_2}) \qquad \text{Equation 10}$$

Part- to Whole-Task Transfer Optimization. For this optimization, the question is whether to practice only single KC components of a whole skill (a conjunctive skill containing at least 2 KCs), only the whole skill or some mixture of the two types of practice. Imagine, for example, practicing simple algebra, and consider that a component of the whole task may be knowing the times tables (the low level component). In this case, the question is how much practice should be allocated to times tables practice before doing algebra practice (the high level component). We might expect that either spending no time on times tables or no time on algebra would likely result in poorer algebra performance than some mixture of these extremes, and that an optimal mixture would allow for the best possible algebra performance. Part to whole transfer optimization allows us to determine this optimal mixture.

To compute this optimal mixture, we model the effect of the low level component LEs on the high level component learning rate. To do this, we must create an equation expressing whole task learning as a function of part task learning. Equation 11 (where subscripts w and p refer to whole and part task respectively) captures the notion that we are looking to maximize whole task time (T_w) * learning rate from an optimally spaced LE, which equals the total learning (this method assumes that all practice occurs at the *PSO* optimal point). Here we specify that *PSO* for a conjunctive task is a function of the strength of the whole (dependent) KC and the probability and latency estimates for the part task. By doing this, we have created a new version of the PSO, PSO_w, that depends on the strength of both the part and whole task KCs. At the same time, we are only concerned with the learning of the whole task, so in practice, the t (retention interval) and g (utility gain) terms are not changed from the original PSO. This provides a mechanism whereby the higher probability and lower latency for a practiced part task increases the expected strength of the PSO_w.

$$\max(T_w PSO_w[p_p p_w, q_p + q_w]) \qquad \text{Equation 11}$$

Having this mechanism, we can compute the time needed to train the part task to maximize its effects on whole task learning. In this case, it can be noted that $totaltime\text{-}T_w$ is spent on the part task, with a learning rate of PSO_p; these values, therefore, control p_p (probability correct) and q_p (latency). This allows us to construct Equation 11, which represents total learning as a function of time spent on the whole task, multiplied by the learning rate for the whole task (which, because of the conjunctive response functions in the PSO_w, is itself a function of time spent on the part task multiplied by the part task learning rate). Equation 11 can then be solved for T_w where $T_w \geq 0$ and $T_w \leq totaltime$.

Practice Length Optimization

Practice length optimization determines the optimal duration of a given LE. PLO relies on the fact that KC study for each LE has diminishing marginal returns as a function of time as shown in various studies (Metcalfe & Kornell, 2003; Pavlik Jr., in press). Equation 12 shows how this optimal study duration is found when the total LE weight score (from Equation 4) divided by the time spent studying is maximized. (Equation 12 assumes some minimum study duration greater than 0 to account for fixed costs.)

$$\max\left(\frac{g(1-e^{-v\,studyduration})}{studyduration + fixedcost}\right) \qquad \text{Equation 12}$$

Practice Quantity Optimization

Practice quantity optimization uses probability correct for long-term practice as a utility measure, then determines how many optimally spaced repetitions it takes to reach the point where probability gain per LE is maximal (the practice quantity optimization point is the p_m value when Equation 13 is maximized) for each item being learned of a set of items.

$$\max\left(\frac{p_{m_n}}{n}\right) \qquad \text{Equation 13}$$

Figure 2 graphs Equation 13 for the parameter set in Pavlik Jr. (2005, Experiment 4) where it was found that 11 practices would have been optimal for each KC, as the maximum value of the probability correct/practices curve occurs at 11 repetitions. It is useful to note that the utility function should reflect the nature of our preferences for target knowledge. For example, if the need for one KC is higher than others, then getting it correct has a higher utility.

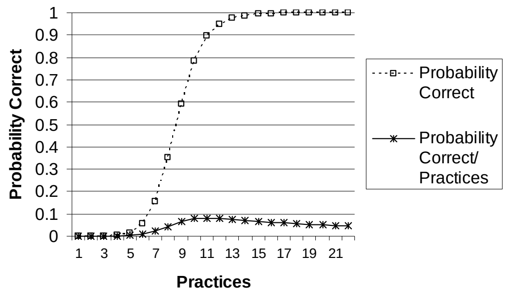

Figure 2. Practice quantity optimization.

To implement this in the model, for instance, we can weight the utility function by the expected frequency of the item we are interested in. This captures the notion that it is twice as important to know a word when that word is used twice as frequently. Having weighted the utility functions, we could then determine a cutoff word frequency below which we will not be concerned with learning the word (this fixes the total amount of time we will need to spend learning the corpus in question).

$$\max\left(\frac{p_{m_n} \cdot expectedfrequency}{n}\right) \qquad \text{Equation 14}$$

Because the weights represent our preferences, other ways of weighting the relative values of different distributions of practice amongst items might further improve the usefulness of such procedures in implementation. For example, items could also be weighted based on the consequences for slow or incorrect performance with the item.

Conclusion

This paper was about a general "microeconomic" method of using a computational model of cognition to compute the efficiency of various decisions that occur during practice. This work is relevant to education because it shows a new approach to understanding how to improve education by considering learning by the student as the measure of profit. In this new approach, the learning of sets of skills can be optimized to maximize output given input.

While we tied this method to an ACT-R cognitive model, there seems no reason why this method could not be used to optimize learning using another computational model. The elegance of the method explained here is that it is theory neutral (given a particular model) and so results in predictions that must be true given the limits of the particular model and the accuracy of the utility function used to capture preferences. In practice, however, the potential of this method can be limited in domains where the complexity of the KC or LEs prevents the clear specification of a utility function to optimize.

Acknowledgments

This research was supported in part by grant from Ronald Zdrojkowski for educational research; the Pittsburgh Science of Learning Center which is funded by the National Science Foundation award number SBE-0354420, and a Graduate Training Grant awarded to Carnegie Mellon University by the Dept. of Education (#R305B040063).

References

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. Journal of Experimental Psychology: Learning, Memory, & Cognition, 23(4), 932-945.

Anderson, J. R., & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. Psychological Science, 2(6), 396-408.

Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning factors analysis - A general method for cognitive model evaluation and improvement. In T.-W. Chan (Ed.), Lecture Notes in Computer Science Intelligent Tutoring Systems (Vol. 4053, pp. 164-175): Springer.

Logan, G. D. (1995). The Weibull distribution, the power law, and the instance theory of automaticity. Psychological Review, 102(4), 751-756.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 102(3), 419-437.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. Journal of Experimental Psychology: General, 132(4), 530-542.

Pavlik Jr., P. I. (2005). The microeconomics of learning: Optimizing paired-associate memory. Dissertation Abstracts International: Section B: The Sciences and Engineering, 66(10-B), 5704.

Pavlik Jr., P. I. (in press). Understanding and applying the dynamics of test practice and study practice. Instructional Science.

Pavlik Jr., P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. Cognitive Science, 29(4), 559-586.

Pavlik Jr., P. I., Presson, N., Dozzi, G., Wu, S.-m., MacWhinney, B., & Koedinger, K. R. (2007). The FaCT (Fact and Concept Training) System: A new tool linking cognitive science with educators. In D. S. McNamara & J. G. Trafton (Eds.). Mahwah, NJ: Lawrence Erlbaum.

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. Journal of Memory and Language, 46(2), 419-440.

Von Neumann, J., & Morgenstern, O. (1944). Theory of games and economic behavior. Princeton,: Princeton university press.