Simulation Models and the Power Law of Learning

Stellan Ohlsson (STELLAN@UIC.EDU)
James J. Jewett (JIMJ@EECS.UMICH.EDU)
University of Illinois at Chicago
Department of Psychology
1007 West Harrison Street, Chicago, IL 60607

Abstract

The power law of learning has frequently been used as a benchmark against which models of skill acquisition should be measured. However, in this paper we show that comparisons between model behavior and the power law phenomenon are uninformative. Qualitatively different assumptions about learning can yield equally good fit to the power law. Also, parameter variations can transform a model with very good fit into a model with bad fit. Empirical tests of learning theories require both comparative evaluation of alternative theories and sensitivity analyses, simulation experiments designed to reveal the region of parameter space within which the model successfully reproduces the empirical phenomenon. Abstract simulation models are better suited for these purposes than either symbolic or connectionist models.

Evaluating Models of Learning

Since the seminal papers by Anderson, Kline and Beasley (1979) and Anzai and Simon (1979), the study of complex learning has seen an unprecedented explosion of theory (Klahr, Langley & Neches, 1987; Chipman & Meyrowitz, 1993). A large number of computational processes with the power to change and improve a knowledge base have been proposed, covering a range of learning scenarios from skill acquisition (Anderson, 1993; Ohlsson, 1996) to conceptual change (Giere, 1992; Thagard, 1992).

The ratio of theory to data is now so high in this field that further progress is dependent upon finding systematic methodologies for evaluating the various theoretical models. Because quantitative regularities are rare in psychology, the field has followed Newell and Rosenbloom's (1981) lead in using the so-called power law of learning as an explanatory target for learning models.

The power law phenomenon consists in the fact that when a learner's performance (measured in terms of the time to complete a practice task or the numbers of errors made per trial) is plotted as a function of trials, the result is a negatively accelerated curve which, moreover, conforms to the shape described by a power law equation. The diagnostic hallmark of a power law curve is that it appears as a straight line when plotted with logarithmic coordinates (as opposed to other types of negatively accelerated curves, e. g., exponential curves, which do not have this feature).

Because the power law phenomenon is quantitatively precise, it appears to be a particularly powerful test of models of learning. Indeed, it "has been accepted as a nearly universal description of skill acquisition to such an extent that it is treated as a law, a benchmark prediction that theories of skill acquisition must make to be serious contenders" (Logan, 1988, p. 495).

Attempts to use the power law of learning to evaluate a learning model usually takes the following simple form: The model is run and its learning curve plotted in log-log space; if the resulting curve is a power law, i.e., if it appears as a straight line, the model is considered validated.

In this paper, we argue that this methodology is too weak to be informative, for two reasons. First, goodness of fit between a theory and an empirical phenomenon is not in and of itself particularly revealing. No theory ever accounts for data completely or precisely. The issue is thus not whether a theory can account for a phenomenon but whether it accounts for the phenomenon better or worse than another theory. Empirical validation must take the form of a *comparative evaluation* in which multiple models are compared with respect to how well they account for the relevant data (Cooper, Fox, Farringdon & Shallice, 1996).

Second, any learning model has parameters (e. g., capacities, constants, thresholds). The possible values of all the relevant parameters define a quantitative space called the *parameter space* (for that model). No model accounts for an empirical regularity (with equal precision) at every point in its parameter space. It is always possible to assign values to the relevant parameters in such a way as to deflect the model's behavior away from the empirical regularity that is the target of the modeling effort. Empirical validation attempts should provide information about how sensitive the model is to such parameter variations, i.e., within which region of parameter space the model accounts for the target phenomenon.

The results of such *sensitivity analyses* (Schneider, 1988) are interesting from two points of view. First, because the parameters might themselves be interpretable in psychological terms, their values might be testable against data. Second, the width of the relevant region is an indicator of robustness. If the region within which the model is successful is narrow, the model's explanation for the phenomenon is not robust. If the empirical phenomenon is robust, this outcome ought to count against the model.

In the following, the need for comparative evaluation and sensitive analyses is illustrated in a series of simulations of the power law of learning. It turns out that both success-driven and failure-driven learning can account for the power law of learning, and so can a mixed model. Explorations of the parameter space show that the degree of fit is sensitive to some parameters but not to others. These results required extensive simulation experiments that would have been difficult to carry out with either a symbolic or a connectionist model, but were relatively easy to do with the *abstract computer model* that we used. We conclude that strict adherence to the sufficiency criterion originally proposed by Newell, Shaw and Simon (1958) can be an obstacle to cognitive modeling.

Basic Model and Method

All simulations were done with one and the same basic model. Each simulation experiment conformed closely to the procedure used in empirical studies of skill acquisition.

Basic Model

The basic model has two components: The task environment and the performance module.

Task environment The task environment was a modified tree structure with a depth of 20 and a branching factor of 10. The root node is the initial problem state and an arbitrarily chosen terminal node is designated the goal node. Nodes not on the path between the root and goal nodes are labeled as errors. The tree was modified to allow more than one path to the goal by the addition of extra links between branches in the tree. This *situation tree* captures the structural features of a 20-step task with multiple correct solutions and 10 alternatives in each step. This level of complexity is at or above the complexity of most tasks used in learning experiments with human subjects. The main simplification is the constant branching factor.

Performance module To perform the task represented by the situation tree is to traverse the tree once, starting in the root node and ending in the goal node. The performance module processes each node by (a) retrieving all outgoing links, (b) deciding probabilistically which link to traverse, and (c) traversing the selected link to the next node.

All links have strengths. Strengths are initially set to unity. The probability p of choosing link L is a function of the current strength s of L. The decision rule is that p is proportional to s. This rule was implemented with an algorithm that multiplies s with a random number between 0 and 1 and chooses the link with the highest product.

Method

The power law of learning is not a behavior. It is a statistical entity constructed by applying operations on raw data. In particular, empirical learning curves are typically constructed by (a) letting several human subjects learn the target task until some criterion of mastery has been reached, (b) averaging performance measures across subjects but within trials, and (c) plotting the resulting averages as a function of trials. In simulating the learning curve, this procedure should be followed as closely as possible

All simulations reported in this paper were run in the following way. Each simulated subject was run until a criterion of three consecutive error-free trials was met. Each simulation experiment consists of 20 simulated subjects.

Time to complete the task is the most commonly used empirical measure in research on skill acquisition. The number of steps to solution is reported here as the closest model equivalent.

Comparative evaluations were accomplished by adding or deleting learning mechanisms to the basic model. Sensitivity experiments were done by varying parameters.

Comparative Evaluation

We compared success-driven and failure-driven learning with each other and with a mixed model that learns both from success and failure.

Success-Driven Learning

The basic model is initialized with a strength of 1.0 on all options. This is equivalent to saying that the model does not know what to do in any of the problem states. Success-driven learning was implemented by incrementing the strength s of a link L with a constant amount ds if traversal of L does not encounter an error signal.

Figure 1 shows the result of success-driven learning when the ds parameter is (arbitrarily) set to .20. The results are plotted with logarithmic coordinates. The data points represent the output from the simulation and the line is the best-fitting straight line through those points.

As a simulation of the power law of learning, the correspondence to human data is good. The fit to a power law is near-perfect and the learning parameter is 1.027, which is close to the values observed in empirical studies (Lane, 1987; Newell & Rosenbloom, 1981).

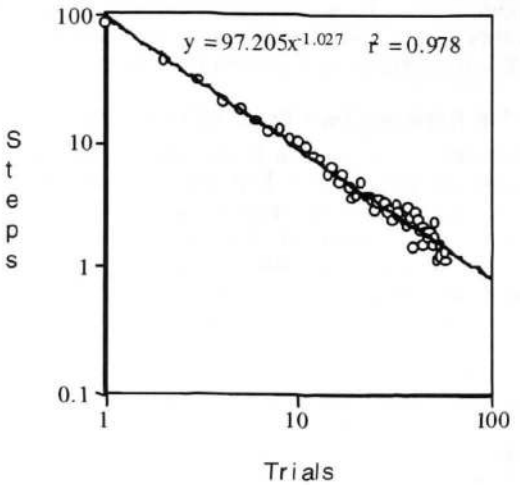

Figure 1. Learning curve for success-driven learning.

In the standard methodology for comparing models to data, the outcome exhibited in Figure 1 would score as a success in accounting for the power law of learning.

Failure-Driven Learning

Our failure-driven learning mechanism decrements the strength of a link L by multiplying it with a constant proportion *fs* if traversal of L encounters an error signal.

Figure 2 shows the result of failure-driven learning when the *fs* parameter is (arbitrarily) set to .50, plotted with logarithmic coordinates. The data points represent the output of the simulation and the line is the best-fitting straight line through those points.

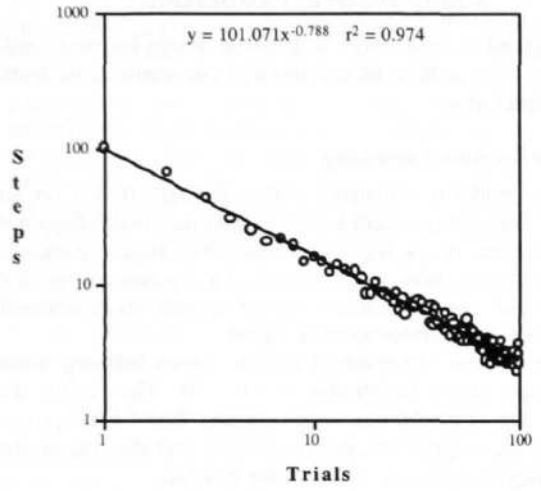

Figure 2. Learning curve for failure-driven learning.

The correspondence to the empirical phenomenon is once again good. The points cluster along a straight line and the learning parameter is .788, which is well within the range of values observed in empirical studies (Lane, 1987; Newell & Rosenbloom, 1981). As a comparison between Figures 1 and 2 shows, the r^2 measure of fit is very similar for success-driven and failure-driven learning (.978 versus .974).

Interaction Between Two Types of Learning

It is reasonable to assume that human beings learn from both success and failure, from both positive and negative feedback. A mixed model was implemented with a success-driven learning mechanism that was identical to the one described above. The failure-driven learning mechanism decremented the strength *s* of a link L by subtracting a constant amount *fs* from *s* if traversal of L encounters an error signal.

Figure 3 shows the result of the mixed model when the *ds* parameter is (arbitrarily) set to .20 and the *fs* parameter is (arbitrarily) set to the same number. Once again, logarithmic coordinates are used. As before, the data points represent the output of the simulation and the curve is the best-fitting straight line through those points.

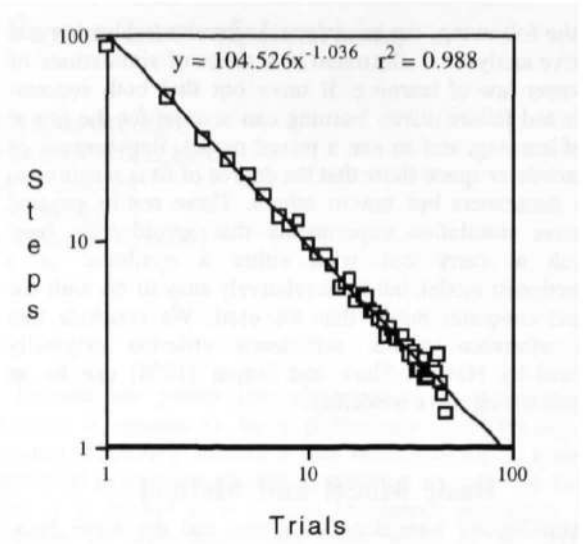

Figure 3. Learning curve for mixed model.

The mixed model also produces a good fit to the power law phenomenon, both in terms of the shape of the curve and the value of the learning rate parameter. Hence, the interaction between success-driven and failure-driven learning is as good an explanation for the observed regularity as either type of learning by itself.

Sensitivity Analyses

In which regions of the parameter space do the above results hold? An exhaustive answer would require a large number of simulation experiments. The four experiments reported below varied the *ds* parameter in the success-driven model, the efficiency of learning, task complexity and the decrementing algorithm in the failure-driven model.

Varying the *ds* parameter Figure 4 shows the result of running the success-driven model with *ds* = .01 instead of .20. The result is clear: The fit to the power law goes away. The data points no longer cluster along a straight line and instead show a distinct curvature.

Additional experiments showed that increasing the *ds* parameter to absurdly large values (e.g., *ds* = 3.0) has a similar effect, bending the scatter plot away from power law fit in the opposite direction. *The region in parameter space for which success-driven learning produces power law learning is narrow.* Because *ds* is a likely individual difference parameter, this finding implies that degree of fit to a power law ought to vary across individuals in learning environments that provide positive feedback. This is a novel prediction.

Varying probability of learning People are not always alert and learners do not always have the ability to react correctly to feedback (positive or negative) during learning. We can model these facts to a first approximation by adding a parameter *lp* that stands for the probability that a learner will learn, given the opportunity. Figure 5 shows the result of running the success-driven model with *pl* = .50, i. e., with the assumption that the learner makes effective use of feedback on half the occasions on which it is available.

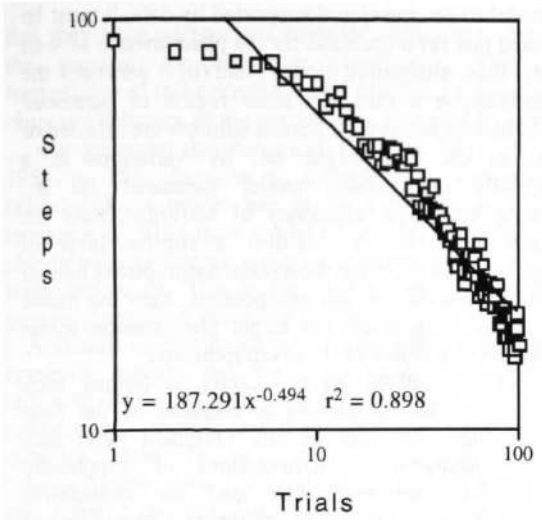

Figure 4. Success-driven learning with *ds*=.01.

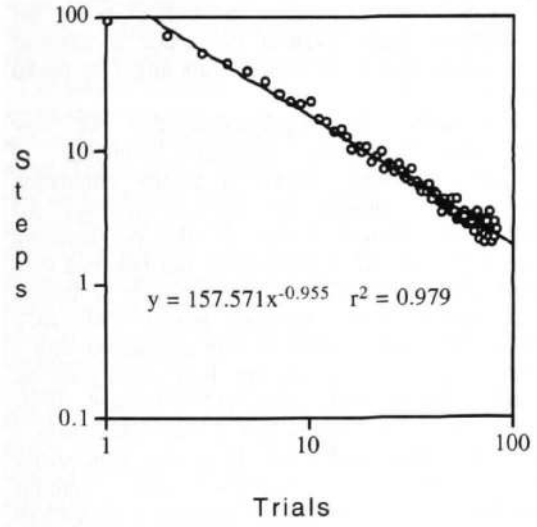

Figure 5. Success-driven learning with *lp* = .50.

As Figure 5 shows, the fit to the power law is unaffected by halving the learning efficiency parameter. Our explorations have shown that success-driven learning produces power law learning curves across a wide range of values for this parameter. Thus, not every parameter affects power law fit.

Varying task complexity It has been suggested that power law learning is exponential learning slowed down by some type of 'mental friction' (Newell & Rosenbloom, 1981). For example, one might expect an increase in task complexity to slow down learning. Figure 6 shows the result of running the success-driven model on two problems with different complexity. The branching factor was 17 in both problems, but the solution path was 10 steps long in one problem and 40 steps in the other. As Figure 6 shows, length of the solution path had no effect on the shape of the learning curve, although it did cause a downward displacement of the curve. Variations in task complexity does not affect the model's ability to account for the power law of learning. This is a strength, because empirical power laws have been obtained in tasks of such widely varying cognitive complexity as pattern recognition (Seibel, 1963) and book writing (Ohlsson, 1992).

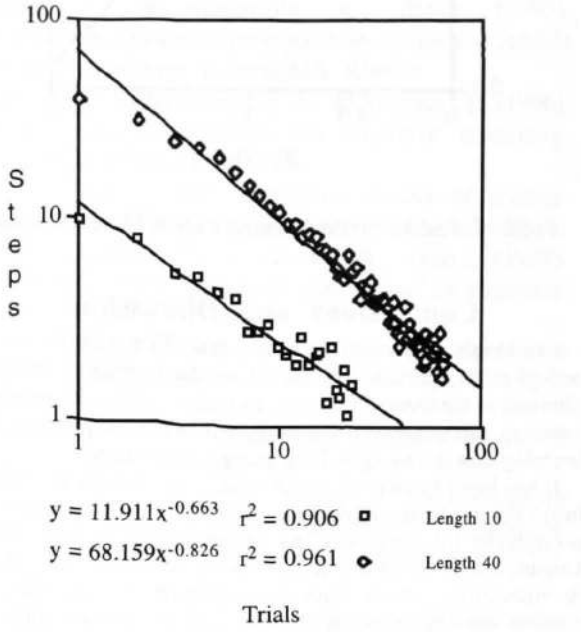

Figure 6. Effect of path length on success-driven learning.

Varying the decrementing mechanism Theoretical principles such as success-driven and failure-driven learning do not uniquely specify the models that we use to test them. As Cooper et al. (1996) have argued, to evaluate the theory underpinning a model we need to vary not only quantitative parameters but also those components of the model that are underspecified by the theory. For example, failure-driven learning can be implemented in different ways. Figure 7 shows the result of running the failure-driven model with an additive instead of multiplicative decrementing mechanism. A fixed amount *fs* was subtracted from the strength of a link when an error signal was encountered.

Once again, the fit to a power law goes away. Notice that in Figure 7, the simulation results have been plotted in a log-linear plot instead of the log-log plot used in the Figures 1-6. The results conform closely to a straight line in log-linear space, the hall mark of an exponential curve (as opposed to power law). Hence, the success of failure-driven learning in accounting for the power law phenomenon is essentially dependent on an implementation detail (the exact decrementing rule) that is not specified by the theoretical principle (failure-driven learning) supposedly being tested.

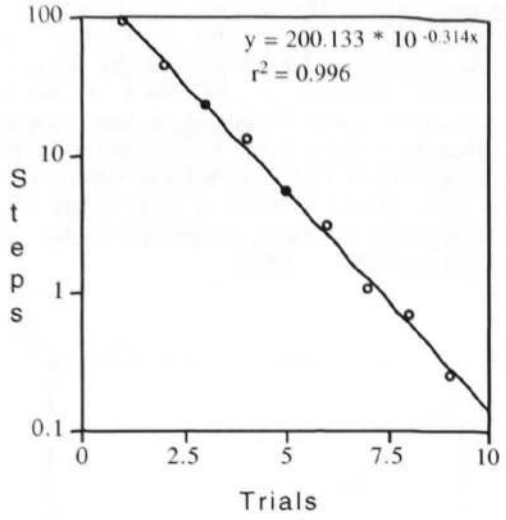

Figure 7. Failure-driven learning with additive decrement.

Conclusions and Discussion

Our results show that the power law of learning is not, in and of itself, particularly useful in discriminating between alternative theories of learning. Indeed, it cannot discriminate between success-driven learning, failure-driven learning and learning that draws upon both success and failure.

It has been known for some time that alternative models, build on qualitatively different assumptions can succeed in accounting for the power law of learning (Anderson, 1993; Logan, 1988; Newell & Rosenbloom, 1981; Shrager, Hogg & Huberman, 1988). However, comparisons between past models were obscured by the fact that the models differ in many respects *other* than their assumptions about learning. The comparisons presented here are more decisive, because the models are exactly alike--indeed, identical--in all other respects than their learning assumptions.

The results of the sensitivity analyses go further. They show that implementation details (quantitative values of parameters, exact algorithms used) that are not specified or determined by the psychological hypotheses underpinning a model can determine whether a model succeeds or fails in accounting for an empirical regularity like the learning curve. In other words, whether a theory is regarded as true or false--as accounting for, or not accounting for, the relevant data--can depend on technical decisions that have little to do with the content of the theory, a long-standing concern with computer simulation (Ohlsson, 1988) recently emphasized by Cooper et al. (1996).

The implication is that *successful tests of simulation models against the power law of learning are meaningless, unless appended with information about the parameter region within which the positive outcome holds.* In general, the standard procedure of running (some version of) a cognitive model and comparing its behavior with human behavior generates little information about the psychological validity of the assumptions and hypotheses that informed the design of the model.

For a model to be considered supported by data, it must be demonstrated that (a) it accounts for the phenomenon as well as, or better than, alternative models, and (b) it generates the target phenomenon within the same region of parameter space as do human subjects. If human subjects are affected or unaffected, as the case might be, by variations in a psychologically interpretable model parameter (e. g., strengthening increment, efficiency of learning), then the model must be shown to exhibit a similar level of robustness. In addition, if the theoretical assumptions behind the model underspecify certain components, then the model must be shown to generate the target phenomenon across different implementations of those components.

Cooper et al. (1996) have recently performed such component variations on an implementation of the Soar model. Consistent with the results presented here, they found that alternative implementations of supposedly innocent model components can alter the quantitative predictions of Soar (and hence its validity as a psychological model, according to the standard canon of model testing).

This extension of the simulation methodology imposes unfamiliar and perhaps unwelcome requirements on cognitive modeling. The amount of work involved in validating a simulation model increases. In principle, a simulation experiment is required for each point in the relevant parameter space. Even if the space is explored selectively, a large number of experiments might be needed to test a model.

More importantly, the requirements that we vary parameters and implement alternative versions of underspecified components require a robust simulation technology. Brittle models are difficult to vary and brittleness is an inherent feature of the programming methodologies that cognitive psychology has inherited from Artificial Intelligence (A. I.). For example, it is unclear whether a success-driven learning model like ACT (Anderson, 1993) would work if it were augmented with a mechanism for learning from failure. It is equally unclear how a failure-driven model like HS (Ohlsson, 1996; Ohlsson & Rees, 1991) would fare if augmented with a success-driven learning mechanism. In general, comparative evaluations, sensitivity analyses and multiple implementations of underspecified components are difficult and cumbersome to carry out with the standard A. I. programming techniques that until recently were the main tools for computer simulation of human cognition.

Cooper et al. (1996) have responded to this situation by proposing that cognitive models should be stated in *executable specification languages.* Such a language is characterized by a high level of abstraction in the statement of a model. In addition, the particular specification language they describe, called *Sceptic*, allows the user to syntactically distinguish between theoretical principles and implementation details.

Although Cooper et al. convincingly demonstrate that an executable specification language like Sceptic is a useful tool, we believe that their response does not go far enough. The problem of distinguishing between theoretical principles and implementation details hides a deeper point. The reason why psychologists have this problem in the first place is

that they have adopted the so-called *sufficiency criterion* as the standard for computer simulation. The original formulation of this criterion stated that "an explanation of an observed behavior of the organism is provided by a program ... that generates this behavior" (Newell, Shaw & Simon, 1958, p. 151). In order to constitute an explanation for behavior X, a model has to be sufficient to mimic or reproduce X. That is, a model of problem solving must be able to solve problems; a model of learning must be able to learn; and so on. This is the explanatory standard that has governed cognitive modeling to date.

Although connectionist models are often contrasted with symbolic models, the former also adopt the sufficiency criterion. Some of the strongest arguments in favor of connectionist models are their ability to perform certain complex tasks (e. g., speech recognition, typing). Connectionists strive to produce intelligent programs. In this respect, connectionist modeling does not differ from symbolic modeling.

However, the sufficiency criterion is not followed in other sciences. Models in meteorology do not produce rain; cosmological models do duplicate the Big Bang; simulations of the economy do not perform monetary transactions; and so on. Such examples suggest that the sufficiency criterion confuses medium and message. A map is not a territory, so why must a model of mind be a mind?

The models discussed in this paper are useful tools for comparative evaluation and sensitivity analyses precisely because they do not attempt to duplicate the processes they model. These *abstract models* are not A. I. programs; they have no knowledge base and no intelligence. These models do what models in other sciences do: They capture as simply as possible general properties of certain formally defined classes of systems (success-driven and failure-driven adaptive agents), and they generate the quantitative implications of those assumptions with respect to observable events and variables; that is all.

Three observations are pertinent here: First, abstract models should be distinguished from the so-called minimal models of the past, i. e., models that attempt to explain behavior in particular experimental paradigms. A unified theory of cognition can be implemented as an abstract model. Second, our models are abstract in a different sense than models stated in the specification languages advocated by Cooper et al. (1996). In our case, low level implementation details are missing, precisely because the models do not carry out the processes they model. Third, we are not arguing that sufficiency models should be abandoned. The more tools we have in our tool kit, the better we can perform the job of researching human cognition.

In conclusion, our experience indicates that abstract models might reveal properties of cognition that are unlikely to be uncovered by models that attempt to satisfy the sufficiency criterion, whether the latter be connectionist or symbolic.

Acknowledgments

Preparation of this paper was supported by Grant No. N00014-95-1-0748 from the Cognitive Science Program of the Office of Naval Research (ONR).

References

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., Kline, P. J., & Beasley, C. M., Jr. (1979). A general learning theory and its application to schema abstraction. In G. H. Bower, (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 13, pp. 277-318). New York, NY: Academic Press.

Anzai, Y., & Simon, H. A. (1979) The theory of learning by doing. *Psychological Review, 86*, 124-140.

Chipman, S., & Meyrowitz, A., (Eds.), (1993). *Foundations of knowledge acquisition: Cognitive models of complex learning*. Boston, MA: Kluwer.

Cooper, R., Fox, J., Farringdon, J., & Shallice, T. (1996). A systematic methodology for cognitive modelling. *Artificial Intelligence, 85*, 3-44.

Giere, R., (Ed.), (1992). *Cognitive models of science*. Minneapolis, Minnesota: University of Minnesota Press.

Klahr, D., Langley, P., & Neches, R., (Eds.), (1987). *Production system models of learning and development*. Cambridge, MA: MIT Press.

Lane, N. (1987). *Skill acquisition rates and patterns: Issues and training implications*. New York, NY: Springer-Verlag.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492-527.

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. Anderson, (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review, 65*, 151-166).

Ohlsson, S. (1988) Computer simulation and its impact on educational research and practice. *International Journal of Education, 12*, 5-34.

Ohlsson, S. (1992) The learning curve for writing books: Evidence from Professor Asimov. *Psychological Science, 3(6)*, 380-382.

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review, 103*, 241-262.

Ohlsson, S., & Rees, E. (1991). The function of conceptual understanding in the learning of arithmetic procedures. *Cognition and Instruction, 8*, 103-179.

Schneider, W. (1988). Sensitivity analysis in connectionist modeling. *Behavior Research Methods, Instruments, & Computers, 20*, 282-288.

Seibel, R. (1963). Discrimination reaction time for a 1,023 alternative task. *Journal of Experimental Psychology, 66*, 215-226.

Shrager, J., Hogg, T., & Huberman, B. A. (1988). A graph-dynamic model of the power law of practice and the problem-solving fan-effect. *Science, 242*, 414-416.

Thagard, P. (1992). *Conceptual revolutions*. Princeton, NJ: Princeton University Press.