From *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society*. Vancouver: Lawrence Erlbaum (1999).

Concrete and Abstract Models of Category Learning

Pat Langley[1] (LANGLEY@ISLE.ORG)
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306 USA

Abstract

In this paper, we compare the rhetoric that sometimes appears in the literature on computational models of category learning with the growing evidence that different theoretical paradigms typically produce similar results. In response, we suggest that concrete computational models, which currently dominate the field, may be less useful than simulations that operate at a more abstract level. We illustrate this point with an abstract simulation that explains a challenging phenomenon in the area of category learning – the effect of consistent contrasts – and we conclude with some general observations about such abstract models.

Introduction and Overview

Learning is one of the ubiquitous aspects of human behavior, so it seems natural that the process of learning has drawn significant attention within both cognitive psychology and artificial intelligence. Over time, different candidate mechanisms have arisen to account for learning phenomena, leading to distinct theoretical camps that have direct analogues across the two disciplines. Another clear parallel lies in the rhetorical stances often taken by authors, which assume that the success of a learning method on a specific problem derives from that method's distinguishing features, rather than from other factors.

In this paper, we review five main paradigms in the computational study of learning, and we consider the mounting evidence that, for purposes of both artifact construction and psychological modeling, these different frameworks typically give equivalent results. Indeed, analysis of successful applications and successful models suggests decisions about how to cast the learning task and how to encode training data are the main source of power in computational learning. This observation leads us to question the usefulness of developing detailed, concrete computational models of human learning.

In response, we draw on the notion of an *abstract* computational model that makes predictions about behavior but that does not actually carry out the task. We discuss some earlier work in this alternative framework that has focused on skill learning, then apply the approach to a phenomenon from category learning – the effect of consistent contrasts – which poses challenges to most computational accounts. We show that a certain abstract model explains this finding without taking a position on the details of representation or learning, whereas another abstract simulation, which matches the assumptions of most concrete models, does not explain the phenomenon. We close with responses to some natural criticisms of abstract models and with comments on their long-term role in developing theories of human behavior.

Rhetoric and Reality in Learning

Much of the research on mechanisms of learning, both within AI and cognitive psychology, has focused on the acquisition of knowledge for classification or categorization. The performance task here involves assigning a new instance or stimulus, typically described using attribute-value pairs, to some category or class, given a known set of mutually exclusive classes. The associated learning task involves finding some function or mapping that categorizes novel instances, given a set of training instances and their assigned classes. The typical performance measure is classification accuracy or error, though measures of speed and typicality sometimes appear as well.

The machine learning community has explored five main representations of knowledge about categories, each which its associated mechanisms. The first major paradigm represents knowledge as *decision lists*, which consist of rules that specify the logical conditions for membership in a category, typically learned one at a time. A second framework represents category knowledge as a *decision tree* that is acquired through a process of recursive partitioning. A third paradigm represents knowledge as a multilayer *neural network*, often relying on a weight-adjusting method known as *backpropagation*. Yet another framework encodes knowledge about categories as experiences or stimuli stored in long-term memory, using *nearest neighbor* or *case-based* methods for classification. A final paradigm uses training instances to update *probabilistic* descriptions, often using simple methods like naive Bayesian classifiers for categorization.

[1] Also affiliated with the DaimlerChrysler Research & Technology Center, Palo Alto, and the Center for the Study of Language and Information at Stanford University.

Superficially, these five paradigms appear quite distinct, and early research in machine learning emphasized differences among them. For example, for many years the common wisdom posited that methods for decision-tree and rule induction were most appropriate for 'symbolic' domains, whereas backpropagation in neural networks was best suited for sensori-motor tasks. Indeed, some felt that such different representations, performance elements, and learning algorithms could not even operate in the same domains. These beliefs were encouraged by the different notations used in various communities, but they were also aided by rhetorical claims, unbacked by evidence, coming from the various camps.

This perception started to change with the first experimental comparisons among different methods for classification learning (e.g., Mooney, Shavlik, Towell, & Gove, 1989). These studies and ensuing ones showed that induction algorithms from separate frameworks, although superficially very different, could operate on the same problems. Their experimental results also suggested that no one induction method was always superior to others, and a decade of experimental comparisons has supported these early results. Although methods for classification learning have steadily improved over time, no one *paradigm* has emerged as superior to others in terms of classification accuracy.

However, contributors to each paradigm have found some quite different factors that affect the success of learning. These include decisions about the formulation of the learning task, the representation or encoding of the stimuli, and the quality of the training cases. Both experimental studies and application efforts suggest that such factors are more important determinants of learning effectiveness than the induction algorithm or the representational formalism itself, although authors seldom emphasize these issues in papers. Langley and Simon (1995) argue that these items – problem formulation, representation engineering, and data collection – are the main sources of explanatory power in machine learning.

Each paradigm in machine learning has a direct analogue in theories of human learning. Techniques for learning decision lists bear a close relation to production-system models of human category learning (e.g., Anderson & Kline, 1979), whereas methods for decision-tree induction are quite similar to psychological models of learning that construct *discrimination networks* (e.g., Richman & Simon, 1989). Backpropagation and its relatives have been used not only for applied problems but also play a role in many models of human learning (e.g., Gluck & Bower, 1988). Case-based methods figure prominently in the papers on human concept learning, where they are known as *exemplar* models (Smith & Medin, 1981), and probabilistic methods have also been proposed as models of human category formation (e.g., Anderson, 1991; Fisher & Langley, 1990).

The literature on computational models of human learning has also seen a period dominated by rhetorical claims. The typical research paper begins by arguing the strengths of connectionism, production systems, or exemplar models, whichever happens to represent the author's paradigm. The text then reviews some psychological phenomena and describes a computational model, cast within this paradigm, that replicates those findings. In closing, the authors conclude that these positive results are evidence for their theoretical framework, ignoring the possibility that the source of explanatory power lies elsewhere, such as in carefully selected stimulus encodings or in a well-crafted training regimen.

The reason for drawing such hasty conclusions are understandable even if the conclusions themselves are questionable. One simply cannot construct a detailed computer simulation of human behavior without making many assumptions, such as representational decisions, that are not central to one's theoretical claims. Naturally, many scientists are tempted to conclude that, when their simulation succeeds at modeling some phenomenon, their core assumptions are responsible rather than the peripheral ones.

Yet not all authors follow this natural inclination, with one revealing counterexample coming from Richman and Simon (1989). They suggest that two alternative accounts of word-recognition findings – connectionist models (which posit parallel processing) and discrimination networks (which posit sequential processing) – are not due to these paradigms' core assumptions. Rather, they argue that a hierarchical representation of words, an auxiliary assumption that both classes of model share, constitute the real source of explanatory power in this domain. We believe many similar examples exist in the literature on computational models of human learning.

Abstract Models of Learning

These observations suggest that traditional computer simulations of human learning, although useful contributions to artificial intelligence and machine learning, may be unnecessary or even misleading in our attempts to explain psychological phenomena. In place of such *concrete* models, we need process models which operate at some more abstract level that lets us make predictions from the central claims of a theory, without needing an overwhelming number of peripheral assumptions.

Of course, there exists a long tradition of such abstract models of learning within mathematical psychology. But many process accounts developed within this framework have drawbacks of their own, in that they usually make constraining assumptions and embody simple theories for the sake of analytical tractability. Such restrictions on analytic models were originally an important factor in the development of computer simulations that actually carry out the task at hand.

However, the decision to work at an abstract level does not mean one must develop an analytic mathematical model; nor does the use of computer simulation mean one's program must accomplish a complete task. Instead, a process-oriented psychologist can develop an *abstract computational model*, a notion championed by Ohlsson and Jewett (1997). In this framework, the scientist still implements a running computer program that predicts behavior, but the system omits details that are not essential to the phenomena it aims to explain. For example, to model learning in problem-solving domains, they retain the idea of search through a problem space, but remove details about the states and operators that define the space. Instead, they specify the structure or connectivity of the space and model the learning process using mechanisms that alter the probability of taking given branches in the future.

Ohlsson and Jewett's goal was to model the power law of learning, in which the rate of improvement decreases with the number of training steps. Simulations on synthetic problems revealed that two learning schemes, involving positive feedback for selecting good branches and negative feedback for bad selections, produced power curves across a broad range of parameter settings. For instance, varying the branching factor, the length of solution, and the probability of feedback did not affect the shape of the learning curve, but extreme parameter settings for success-driven learning gave different simulated behavior. Moreover, failure-driven models that incorporated additive weight reductions in response to negative feedback produced exponential curves, although multiplicative updates gave the power law.

Another abstract computational model of learning comes from Rosenbloom and Newell (1987), who also focused on the power law. Their primary aim was to develop a concrete computer simulation that exhibited this effect on a finger-manipulation task. The key idea in their model is that humans acquire *chunks* which let them link complex perceptual configurations to complex actions, thus reducing the need to carry out multiple reasoning steps at the cognitive level. Rosenbloom and Newell embedded their learning mechanism in a detailed theory of the human cognitive architecture, cast as a production system, and showed that their mechanism for chunk acquisition reduced response time with practice. However, to actually fit the psychological data, they invoked a simple abstract model with four parameters that embodied the core assumptions of their chunking theory.

Shrager, Hogg, and Huberman (1988) present yet another explanation of power-law learning. Like Ohlsson and Jewett's, their abstract model describes a problem space only in terms of nodes and links, along with the probability that a selected branch will lead toward the goal node. Their computer simulations show that power-level behavior can result from two quite different learning processes. One mechanism (similar to Rosenbloom and Newell's) creates new links from a problem's initial state to its goal state, letting the problem solver make future traversals in one step. Another mechanism (closer to Ohlsson and Jewett's) alters the probability of traversing a link based on whether it led to a solution. Shrager et al. also carried out an average-case analysis of their task, which gave good fits to simulated behaviors.

Langley (1996) reports a rather different abstract model for the task of flying an aircraft simulator through a three-dimensional slalom course. His model's central assumptions are that differences among subjects are due to differences in sensing skills, and that the main form of learning involves improving the ability to focus on relevant features during skill execution. Langley describes an implementation of this abstract model of sensory learning, along with a system that searches the space of parameter settings in order to fit the model to the experimental data. He compares the sensory-learning framework to an alternative model based on the power law, finding that the latter fits the data slightly better but that it requires many more parameters.

There are clear kinships between these abstract simulations and models from mathematical learning theory, such as Estes' stimulus sampling account of learning. Both frameworks typically assume that subjects' decisions are probabilistic in nature and that learning follows from simple changes to probability distributions. As we have noted, the key difference lies in abstract models' reliance on computer simulation rather than detailed analysis, which supports a wider range of process models. A similar relation holds with respect to the average-case analyses occasionally published in machine learning.

Of course, the different approaches to process modeling are not mutually exclusive. The Rosenbloom and Newell work showed that concrete and abstract simulations can coexist, and the Shrager et al. analysis made the same point with respect to abstract simulations and purely analytical models. Ohlsson and Jewett's contribution was the realization that neither mathematical analysis nor the concrete model are really necessary, and that researchers may often find it useful to work entirely at the level of abstract computational models. Nevertheless, research in this paradigm remains rare, especially in the otherwise well-studied topic of category learning. In the remaining pages, we apply the abstract modeling framework to an intriguing phenomenon in this area.

The Effect of Consistent Contrasts

As we have noted, considerable effort has gone into computational models of human category learning, typically using techniques very similar to those from machine learning. For example, Kruschke's (1992) Alcove incorporates a variant on the nearest-neighbor method that places weights on attributes, Martin and Billman's

Table 1: Schema for the stimuli used in Billman and Dávila's (1995) experimental study of category learning. 'Consistent contrast' subjects saw instances from categories characterized by the same two attributes, whereas those in the inconsistent contrast condition learned categories characterized by different attributes.

	Cons. contrast	Incons. contrast
Category 1	11 xx xx	11 xx xx
Category 2	22 xx xx	xx 22 xx
Category 3	33 xx xx	xx xx 33

(1992) Twilix constructs a form of multivariate decision tree, and Anderson's (1991) RA model bears a close relation to the naive Bayesian classifier. All three systems have shown good matches to experimental results on human category learning. However, here we consider an interesting phenomenon that seems difficult to explain within the standard theoretical frameworks.

Billman and Dávila (1995) noted that most psychological studies of concept induction assume that some attributes are relevant and others irrelevant, but that the *same* ones are relevant to each category. They hypothesized that subjects would find concepts easier to learn when such *consistent contrast* occurs than when distinct categories are defined by *different* features. Table 1 shows the structure of the stimuli Billman and Dávila used to test this hypothesis, using a cover story in which subjects classified animals from an alien zoo and received feedback after each guess. Both conditions involve three classes and six attributes with three values each; moreover, all target concepts involve a conjunction of two relevant features. However, in the consistent condition, the same attributes are relevant to recognizing cases from all three classes, whereas in the inconsistent condition, a different pair plays this role for each class.

The learning curves in Figure 1 (a) show a clear difference between the two experimental conditions. Subjects who dealt with consistent contrasts improved very rapidly, achieving over 90 percent predictive accuracy after only ten training stimuli. Subjects in the inconsistent condition hovered around 50 percent during most of the 45 instances, better than the 33 percent that results from random guesses, but far below the accuracy for the consistent subjects. Separate tests on novel stimuli, some that matched the intended category definitions and others that did not, showed that subjects in the consistent condition were much more accurate at this task as well.

Naturally, Billman and Dávila attempted to explain this phenomenon using existing computational models of category induction. However, simulation runs with Kruschke's Alcove predicted no differences between the two conditions, and similar studies with Anderson's RA indicated a slight advantage for the *inconsistent* condition. Even runs with Martin and Billman's Twilix, which they had expected to reflect the observed differences, failed to produce the desired result. Further analysis suggested that all three models lack a strong bias toward category descriptions incorporating fewer attributes overall, which seems the obvious explanation for the large difference in learning rate.

Of course, we could incorporate such an inductive bias into yet another concrete simulation of category learning, based on one of the above models or embedded in a new one entirely. But this would require us to adopt a position on the representation of knowledge, to select a complete performance element, and to propose a detailed learning algorithm. Yet the above account states that none of these factors are important in explaining the consistent contrast phenomenon. Rather, the key issue is whether learners are biased toward category descriptions that, across concept boundaries, require fewer features. Thus, this seems like an ideal context in which to illustrate the notion of abstract models.

An Abstract Model of Contrast Effects

We want our abstract model to make as few assumptions about representation, performance, and learning as necessary to account for the phenomena at hand. However, we can view all induction methods as constructing decision regions that partition a multi-dimensional space of instances or stimuli. Moreover, all basic induction algorithms incorporate some type of *locality* bias, so they are typically more accurate on test cases that fall near to observed training cases in this space. We would like a modeling framework that reflects this bias without committing to a particular encoding of learned knowledge.

For discrete domains like the one in Billman and Dávila's study, we can model conceptual knowledge as a table with c columns (one per attribute) and r rows, with each row specifying a unique combination of attribute values. We also need one extra column to specify the category or class associated with each value combination or to state that the class is unknown for this situation. This notation lets us describe arbitrary contrasting concepts that map from combinations of discrete values to class labels. Given a attributes with v values each, we can have a table with $a + 1$ columns and v^a rows, but many concepts are much simpler in nature. For example, if only c attributes are relevant, we need only include c columns, which means we need at most v^c rows. And not all possible combinations of values may occur in practice, which lets us reduce the number of rows still further.

Our performance and learning elements are similarly abstract. Given a test stimulus, described as a attributes and their values, we assume the subject finds the table's row whose c attribute values match this instance. If the

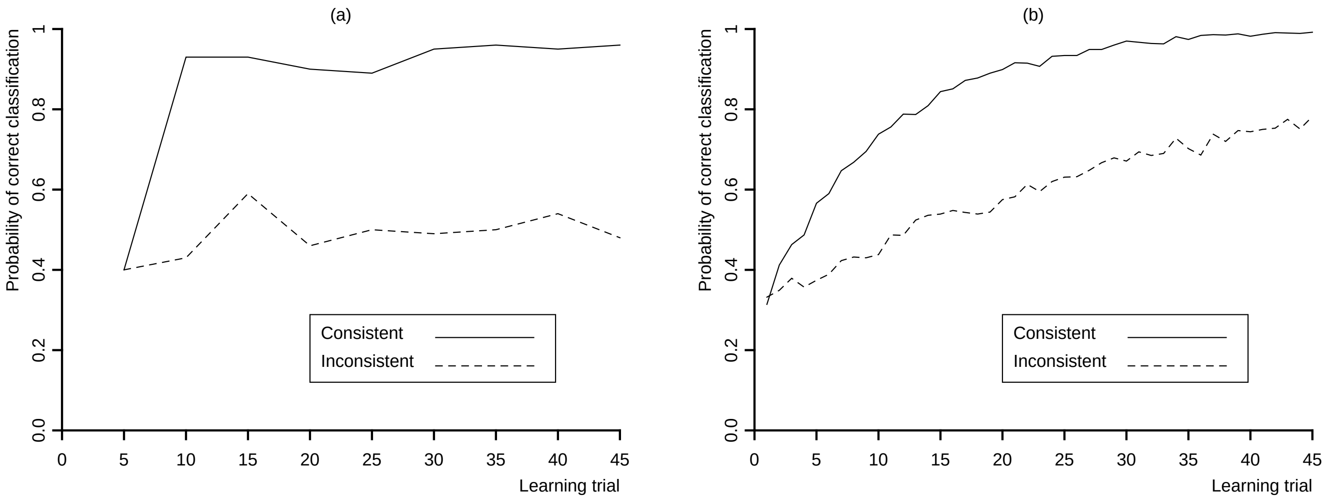

Figure 1: (a) Learning curves that Billman and Dávila's observed for subjects in conditions involving consistent and inconsistent contrast; and (b) learning curves that the abstract model predicts for the same conditions when $p = 0.3$.

row has an associated class, then the subject predicts it; otherwise he selects a class at random from a uniform distribution. We posit two distinct learning mechanisms, one that selects relevant features and another that assigns class labels to rows in the table. We assume that feature selection happens early in the learning process, and thus we model only its result in terms of the number of columns c in the table. For labeling, we assume that each time the subject sees a training instance that matches a given row with an unknown label, he stores, with probability p, the label observed with that stimulus.

When we instantiate this model for Billman and Dávila's two conditions, we see that it should predict quite different behavior. For the situation involving consistent contrast, we have two attributes that are relevant across all categories, giving a table with only two columns. Moreover, since different values on the other four attributes do not matter, we need only three rows in the table, one for each co-occurring pair of relevant values. On the other hand, the table for the inconsistent condition requires six columns, since all six attributes play a role in some concept description; this means we must have 12 rows, one for each combination of values in the training set. Even ignoring the stage of feature selection, which we do not model, subjects should take longer to master categories that require the larger table.

We implemented this abstract model as a simple Lisp program that accepts as input the number of simulated subjects, rows, classes, and training items, along with the probability p of learning on each trial. Figure 1 (b) shows the behaviors that the model generates when we set p to 0.3 and averaged over 1000 simulation runs. As intended, there is a clear difference between simulated subjects under conditions of consistent and inconsistent contrast, with the former learning much more rapidly than the latter. The match to Billman and Dávila's results is only qualitative, as the simulated learning curve for the consistent condition is slower, and the one for the inconsistent condition higher, than they observed.[2] Altering the parameter p does not help, since this speeds or slows the curves for both conditions. However, Billman (personal communication, 1998) reports that using stimuli with different within-class similarity reduces the separation between the two curves. An extended model might incorporate such additional factors, but the current one still produces the basic effect intended.

We can contrast this qualitative behavior with that for a different abstract model that operates in the same manner but that does not include feature selection. We can simulate this situation by assuming that the tables encoding the learned knowledge have the same number of columns and rows for both the consistent and inconsistent conditions. Thus, they predict identical behavior for subjects in both situations. As such, it constitutes an abstract version of the concrete models developed by Martin and Billman, Anderson, and Kruschke. But, to reiterate, we need not descend to their detailed level to explain the consistent contrast effect.

Closing Remarks

In the preceding pages, we reviewed the main research paradigms in machine learning and their links to computational models of human learning. We also argued that, for purposes of both developing artifacts and matching human behavior, one can usually achieve very similar results with each of the various approaches. Moreover, we claimed that the source of explanatory power often

[2] On novel test items, the model also predicts very high accuracy for the consistent situation and chance for the inconsistent condition. In this case, the experimental differences are smaller than the model predicts, but the behaviors again match at the qualitative level.

lies not in whether one uses rule induction, neural networks, exemplar models, decision trees, or probabilistic schemes, but rather in the features used to describe experience, the formulation of the problem, and the nature of the training items. Our response was to recommend the use of abstract computational models to explain phenomena, rather than the concrete models that have direct analogs in machine learning. We reviewed some examples of abstract models and applied this approach to specific experimental results in category learning.

Before closing, we should examine some likely criticisms of abstract models. For example, one might claim that such models merely 'describe' the data rather than explain them. But the models we have reported all posit explicit (although abstract) processes, and thus embody some form of explanatory structure. A more interesting question concerns whether such models' assumptions are necessary or merely sufficient to explain the phenomena. Since we reviewed three abstract models of the power law, each making somewhat different assumptions, they clearly constitute the sufficient variety, but necessity is a difficult hurdle to leap in any science.

A deeper criticism is that, to date, abstract modeling efforts have focused on explaining isolated phenomena. Clearly, we do not want to develop 20 unrelated models of category learning, one for each robust phenomenon in the literature. A more desirable approach would imitate older sciences like physics, which devise separate models for each phenomenon but constrain them with links to deep theoretical principles. The concrete modeling community has made some progress on this front, as in using discrimination networks to explain diverse memory phenomena (H. A. Simon, personal communication, 1998), but the same strategy should work for abstract models.

In the long term, these two frameworks need not remain antithetical. As we gradually extend abstract models to cover more phenomena, we must place ever more constraints on them to ensure consistency with previous accounts. At some point, we may even have enough constraints to take defensible positions on issues like the underlying representation of knowledge, the performance mechanisms that operate on that knowledge, and the learning processes that generate it. Eventually, we may have enough data to justify the construction of concrete models or even a unified theory of the cognitive architecture that covers behavior in many domains. However, we do not feel the study of human learning has reached that stage, and abstract models, even isolated ones that focus on specific results, seem worthy of increased attention.

Acknowledgements

We owe thanks to Dorrit Billman and Michael Pazzani for discussions that led to many of the ideas in this paper, and to Dorrit Billman for making available the results of her experimental study.

References

Anderson, J. R., & Kline, P. J. (1979). A learning system and its psychological implications. *Proceedings of the Sixth International Joint Conference on Artificial Intelligence* (pp. 16–21). Tokyo: Morgan Kaufmann.

Anderson, J. R. (1991). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Billman, D., & Dávila, D. (1995). Consistency is the hobgoblin of human minds: People care but concept learning models do not. *Proceedings of the Seventeenth Conference of the Cognitive Science Society* (pp. 188–193). Pittsburgh: Lawrence Erlbaum.

Fisher, D. H., & Langley, P. (1990). The structure and formation of natural categories. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26). Cambridge, MA: Academic Press.

Gluck, M. A., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, *27*, 166–195.

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

Langley, P. (1996). An abstract computational model of learning selective sensing skills. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 385–390). Lawrence Erlbaum.

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, *38*, November, 55–64.

Martin, J., & Billman, D. (1991). Variability bias and category learning. *Proceedings of the Eighth International Workshop on Machine Learning* (pp. 90–94). Evanston, IL: Morgan Kaufmann.

Mooney, R., Shavlik, S., Towell, G., & Gove, A. (1989). An experimental comparison of symbolic and connectionist learning algorithms. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* (pp. 775–780). Detroit: Morgan Kaufmann.

Ohlsson, S., & Jewett, J. J. (1997). Simulation models and the power law of learning. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 584–589). Stanford, CA: Lawrence Erlbaum.

Richman, H. B., & Simon, H. A. (1989). Context effects in letter perception: Comparison of two models. *Psychological Review*, *96*, 417–432.

Rosenbloom, P. S., & Newell, A. (1987). Learning by chunking: A production system model of practice. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production system models of learning and development*. Cambridge, MA: MIT Press.

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.