Chapter 2

Skill acquisition: History, questions, and theories

Only those who have the patience to do simple things perfectly ever acquire the skill to do difficult things easily.

Author Unknown

2.1 What is skill acquisition?

Skill acquisition is a specific form of learning. For our purposes, it will be sufficient to define learning as the representation of information in memory concerning some environmental or cognitive event. Thus, learning refers to an organism storing something about its past in memory. Skill acquisition refers to a form of prolonged learning about a family of events. Through many pairings of similar stimuli with particular responses, a person can begin to develop knowledge representations of how to respond in certain situations. These representations have some form of privileged status in memory because they can be retrieved more easily and reliably than memories of single events. Thus, skilled behaviours can become routinized and even automatic under some conditions.

The range of behaviours that can be considered to involve skill acquisition could potentially include all responses that are not innate. That is, any response that can be learned can potentially be refined with practice, given the right conditions. Furthermore, these responses are not restricted to overt behaviours. For instance, consider learning to recognize a symbol such as ‘f’ as the letter F, or ‘dog’ as something that denotes ‘a four legged mammal, sometimes referred to as man’s best friend’. In both cases, a mental representation has been developed that evokes some response (in this case the appreciation of meaning) upon encountering a particular stimulus. Furthermore, such recognition is normally not accompanied by any conscious deliberation over meaning. The meaning just pops into our heads. Clearly the same sort of process is operating when experienced chess players know the most appropriate next move, when expert radiologists spot abnormalities in an X-ray slide, and when we recognize others as friends or family. Of course, skilled behaviour also includes overt behaviour. Writing, playing a musical instrument, driving a car, and laying bricks are all examples of behaviours that could not be performed well without a great deal of practice. Thus skilled behaviour and the acquisition of skills is potentially involved in the full range of human behaviour. This is certainly the point of view we have adopted, although it appears to us that it is not an opinion that is held universally. Certainly it is clear that many behaviours, and in particular cognitions, are studied without any consideration of their history. That is, has there been any form of skill acquisition, and if so, how much, and under what conditions? We demonstrate in this book why such questions are important, even in areas where learning seemingly plays no role. In this chapter, we consider some of the history of research in skill acquisition, some of the issues that have been of concern to researchers in this area, and some of the theories that have been proposed as accounts of the major phenomena associated with skill acquisition. In particular, we begin to evaluate the theories with respect to how well they meet the Skill Acquisition Phenomena Criterion introduced in the previous chapter. That is, how well do the theories account for skill acquisition phenomena?

2.2 Great questions and issues

From the late 1800s to the early 1960s, research into skilled performance and skill acquisition was largely devoid of any clear direction, theory, or results. The research was mainly applied in nature and concerned motor skills almost exclusively. The focus was typically on discovering the best methods for training motor skills, where the best methods were those that enabled the fastest learning and the greatest transfer to different situations and tasks. The major areas of research during this period are summarized below. A more comprehensive treatment of this research can be found in Adams (1987).

2.2.1 Plateaus

Research on complex skills really began with the work of Bryan and Harter (1899). They trained subjects in the sending and receiving of Morse code signals and examined the learning curves of these two tasks. The most interesting result of this research was the observation of plateaus in the learning curves of the receiving task. These plateaus represented periods during training where subjects' performance did not improve. However, further training usually resulted in further improvement. Similar plateaus were also reported during the training of typewriting (Book, 1925). The observation of plateaus led Bryan and Harter (1899) to propose that skill learning involved the acquisition of a hierarchy of habits. In a statement that has no doubt influenced some of today's theories of skill acquisition, Bryan and Harter described plateaus as periods where 'lower-order habits are approaching their maximum development, but are not yet sufficiently automatic to leave the attention free to attack the higher-order habits' (p.357).

The concept of plateaus in learning curves has not enjoyed wide support. Since Book's research, there has been little replication of the original findings (Adams, 1987; McGeoch, 1931, 1942). In addition, the extent to which plateaus can constrain theories of skill acquisition is questionable, since any variable that retards learning will produce them (Hunter, 1929). Still, as indicated above, the hierarchical view of skilled behaviour is fundamental to many modern theories of skill acquisition. Moreover, evidence of plateaus in learning curves does appear regularly in the literature, particularly where the tasks being performed involve the acquisition of many skill components that may be acquired at different rates (e.g. Thomas, 1998).

2.2.2 Part–whole training

One of the aims of early research in skill acquisition was to determine whether training on parts of a task and then combining these parts could be more efficient than training on the whole task. The benefits of such a training strategy are obvious, especially where the criterion task involves a large degree of cost or danger. For example, there is obvious value in a training method that would allow a substantial portion of pilot training to be achieved on the ground. Unfortunately, generalizations are not easy to arrive at in this area (McGeoch and Irion, 1952). The best approximation to a definitive answer to this research issue is that the relative benefits of part versus whole training are dependent on the task. Welford (1968) suggests that whole task training is the most efficient means of learning to perform tasks that involve highly interrelated activities, such as flying an aircraft. In contrast, those tasks that involve components which are performed in a fixed order and are largely independent of each other appear to benefit most from part training (Welford, 1968). At present, no theory of skill acquisition purports to account for this relationship between task type and the most efficient means of training.

2.2.3 Massed versus distributed practice

Another research issue in the history of skill acquisition that focussed on the best means of practicing a task concerned the relative benefits of massed and distributed practice. That is, which type of practice is the most efficient training method: continuous practice in one long session (massed) or spaced practice in a number of sessions separated by time intervals of a certain duration (distributed)? All of the research in this area is beset with the problem of defining training efficiency when the cost of training is dependent on both the time spent training on the task and the length of time from the beginning of training to the final testing. Despite this problem, the most popular generalization is that distributed training is the most efficient method (McGeoch, 1931; Welford, 1968). Adams (1987), however, rejected this conclusion. After reviewing 100 years of research on this issue, Adams concluded that distributed practice does not improve learning relative to massed practice, but instead improves the momentary level of performance only. For example, subjects trained with massed practice and then examined under distributed conditions perform equally as well as subjects performing under distributed conditions throughout (Adams and Reynolds, 1954).

The issue has been further complicated by a meta-analysis study reported by Lee and Genovese (1988). After reviewing 116 studies that were directly relevant to a comparison of distributed and massed training, Lee and Genovese found that 47 studies were suitable for a meta-analysis of effect size. On the basis of this analysis, it was concluded that distributed practice led to better acquisition and retention than massed practice. Lee and Genovese suggested also, however, that this generalization may only apply to continuous tasks (e.g. tracking an object on a computer screen with a mouse). This suggestion was supported by the results of a subsequent study (Lee and Genovese, 1989) that also indicated that acquisition and retention of skill on a discrete task (e.g. press a button in response to a stimulus) was better under massed practice conditions.

2.2.4 Knowledge of results

An important concern to some researching skill acquisition was whether performance improves without knowledge of results. Bartlett (1948) stated an answer to this question: practice without knowledge of results does not improve performance. This generalization has been widely accepted, although explanations of this effect, ranging from motivational to associative, are subject to debate (Adams, 1987).

The effects of withdrawing and delaying feedback have been examined extensively and some clear results have emerged. Welford (1968) reports a number of these: (1) When knowledge of results in motor tasks is delayed, learning is slowed. However, movement accuracy on most trials is affected to a small extent only. The slowing of learning is due to a greater proportion of trials that involve relatively larger movement errors. (2) When subjects perform some other activity between a trial and knowledge of the results of that trial, learning is slowed. (3) Learning is slowed as the gap between feedback and the following trial is lengthened. (4) Increasing the precision of feedback increases the accuracy of motor performance. (5) If knowledge of results is provided during training and then subsequently removed, performance deteriorates.

2.2.5 Learning curves

One of the methods researchers have used to track improvement on a skill has been to plot performance as a function of amount of practice. Generally, when performance improves with practice, the graph of improvement is a smooth curve. Such curves are known generically as learning curves. Learning curves first appeared in psychology over a hundred years ago with the work of Ebbinghaus (1885). He was interested in the factors that determined how easily things could be stored in and forgotten from memory. Ebbinghaus was particularly dedicated; rehearsing lists of nonsense syllables (e.g. FOJ) to himself thousands of times to try to commit them to memory. One of the major influences of Ebbinghaus's work on Psychology was the presentation of his results in graphical form to show the number of syllables remembered as a function of the number of repetitions. Typically, memory increased as repetitions increased. The results portrayed on such a graph also had a characteristic shape, with memory improving substantially in the early stages of rehearsal, and improvements getting smaller and smaller as rehearsal continued.

Following Ebbinghaus, learning curves became commonplace in any study of learning. Psychologists of the behaviourist school (e.g. Ferster and Skinner, 1957) were particularly enamoured with learning curves as a means of representing the conditioning of rats and pigeons. Hungry animals would be placed in small boxes fitted with levers (for rats) or buttons (for pigeons). Pressing on these objects would typically result in a reward in the form of a food pellet. Psychologists were interested in the pressing rate exhibited by the animals in response to differing schedules of reward. Typically, the responses would be automatically recorded by a pen on continually moving paper. This resulted in curves that indicated the extent of learning exhibited by the animals.

Although the behaviourist influence on psychology has waned somewhat, learning curves are still used extensively in psychology. They are typically used for representing how the accuracy or speed of performance improves with practice. These curves, particularly those measuring speed, have characteristic shapes. Early in practice, speed improvements are dramatic, but taper off with continued practice—a case of diminishing returns. Snoddy (1926) was the first to note that when the logarithm of performance time is plotted against the logarithm of amount of practice, a straight line typically results. This indicates that performance time can be described as a power function of practice, as indicated by the following equation:

$$T = X + NP^c \tag{2.1}$$

In Equation 2.1, T is the performance time on a task, P is the amount of practice on the task, $X+N$ is performance time on the first trial of the task, and X is the performance time after an infinite amount of practice. Power functions where $X = 0$ often provide very good fits to data, especially where large amounts of practice are involved. The parameter c in Equation 1 is the learning rate. The value of c is less than zero, to match the negative accelerated feature of the learning curve, and is also usually a value between 0 and -1. The closer the value of c is to -1, the faster the learning rate.

Crossman (1959) re-analysed the data of many experiments in varying domains, such as card sorting, addition of digits, and cigar rolling, and noted that learning in all of these tasks conformed to what he referred to as 'de Jong's law'. De Jong (1957) had also noted the power function regularity in learning curves. Newell and Rosenbloom (1981) also examined learning curves in a wide range of domains. They examined the ability of power functions to fit learning curves in comparison with other functions that are also negatively accelerated, such as the exponential function and the hyperbolic function, and concluded that the power function regularly provided a closer fit to the data. Newell and Rosenbloom, like Crossman, were impressed with the apparent lawfulness of the regularity in the learning data, so much so that they referred to the regularity as the power law of practice, and noted that it was one of the few laws in Psychology.

Indeed the presence of power functions in human learning data is so ubiquitous that the power law of practice has almost become an accepted fact in Psychology. As a small taste of this ubiquity, consider the following examples. First, Figure 2.1 demonstrates the typical shape of a power function learning curve. Part (a) contains the reaction time data for a group of people performing a computer task involving simple arithmetic (see Speelman and Kirsner, 2001, for details of the task). Part (b) of the figure presents the same data on log–log axes. In both parts, the line represents the best-fit power function. Similar curves are apparent in the data presented in Figures 2.2, 2.3, and 2.4. This data

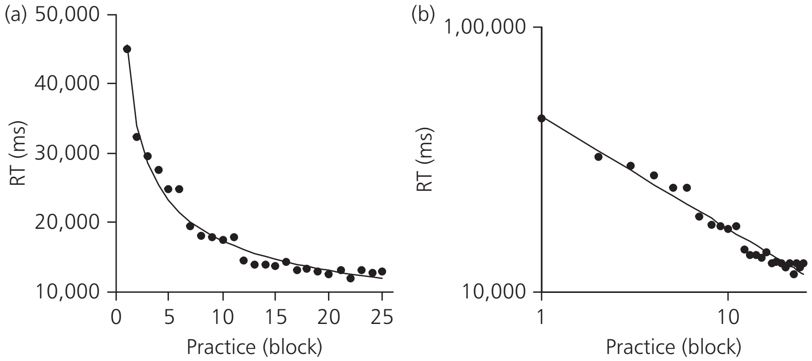

Fig. 2.1 Learning curves on linear–linear axes (a) and log–log axes (b) for people solving simple arithmetic problems (from Speelman and Kirsner, 2001). Lines represent best-fit power functions.

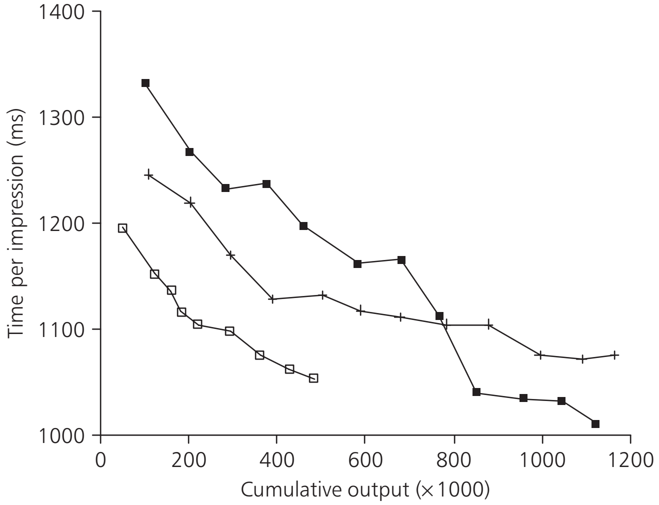

Fig. 2.2 Learning curves of three pressmen, representing time to make an impression in printing as a function of the number of impressions made (from Levy, 1965).

represents performance on a wide range of tasks: making print impressions (Figure 2.2); writing books (Figure 2.3); and even the production of ships (Figure 2.4).

Most discussion of the power law centres on data regarding the speed with which people perform tasks. It is also the case, however, that other measures of performance give rise to power function learning curves. For example, the number of errors committed on a task is typically reduced with practice

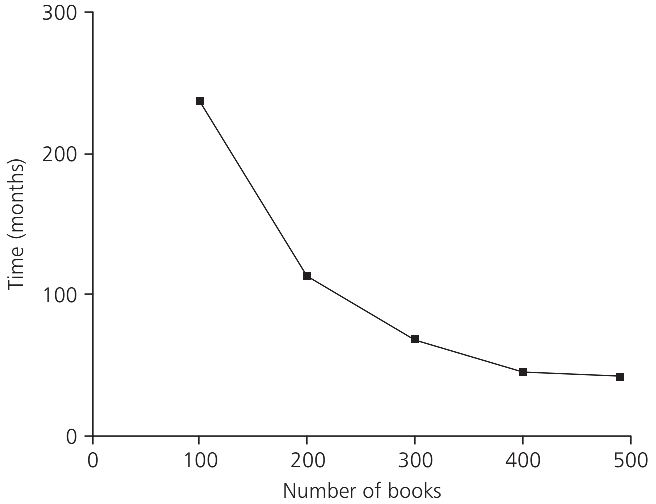

Fig. 2.3 Learning curve for Isaac Asimov, representing time to write a book as a function of number of books written (from Ohlsson, 1992).

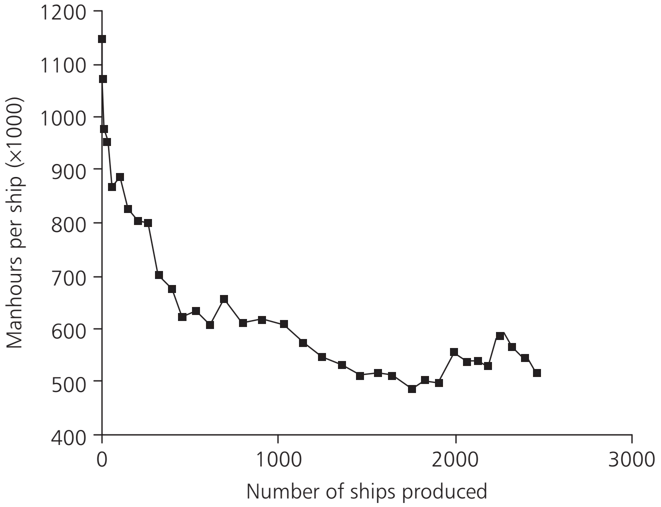

Fig. 2.4 Learning curve for the production of Liberty Vessels, representing time to produce each ship as a function of the number of ships produced (from Searle and Gody, 1945).

(e.g. see Figure 2.5). Whenever the power law is considered, it usually refers to changes in performance time on tasks where error rates are very low, and change little with practice.

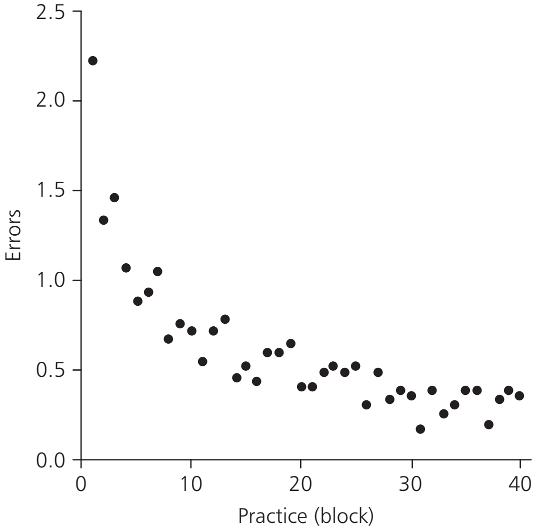

Fig. 2.5 Error rate as a function of practice on a simple arithmetic task (data from an unpublished experiment by Speelman, 1999).

The acceptance of the power law of practice within Psychology has grown to the extent that some researchers (e.g. Anderson, 1983a; Logan, 1988) have suggested that for any theory of skill acquisition to be considered seriously, it must be able to provide an account of the law. Certainly, anyone who has proposed a theory of skill acquisition in the last 25 years has promoted the ability of the theory to account for the power law as an important indicator of the theory's validity. A number of such theories are considered later in this chapter. It is also important to note that there have been suggestions recently that the power law may not be as lawful as has generally been believed. In addition, there is evidence that the shape of learning curves is affected by transfer. These issues are discussed in Chapter 3.

2.2.6 Transfer of training

As with the part versus whole training issue, transfer of training research has typically had an applied focus. The implicit motive in this research is usually to demonstrate that training in one situation will improve performance in another situation, with experimental manipulation aimed at either increasing or decreasing the extent of this transfer. An obvious example is research that looks at the relationship between training in a flight-simulator and subsequent performance in a real aircraft. Historically, the challenge in such research has been to investigate what features in the training and transfer situations determine the amount of transfer that occurs. In other words, what is transferred from one situation to another that can either benefit or impede performance in the second situation? The general aim, then, has been to determine when skills learned in one task can be transferred to performance of another task, with the hope of devising more efficient training methods. Efficiency in this context refers to some notion of the cost benefits of training, where training on a less expensive and time-consuming task may reduce the training time required on the target task. However, an influential theory of the principles underlying transfer—the identical elements theory of Thorndike (Thorndike, 1906; Thorndike and Woodworth, 1901)—suggested that there are no free rides in skill acquisition.

Thorndike's identical elements theory states that transfer between two tasks is determined by the extent to which the tasks share the same content. Thus, the more that is common between tasks in terms of stimuli, responses, or stimulus–response pairs, the greater will be the transfer. The theory has enjoyed widespread support from research in both verbal learning (e.g. McGeoch and Irion, 1952; Osgood, 1949) and motor learning (e.g. Adams, 1987; Duncan, 1958).

The identical elements theory was developed in the early days of behaviourism and so was couched in stimulus–response terms rather than cognitive terms. As a result, criticism was levelled at the theory's inability to account for transfer that apparently was not related to commonalities between tasks but was associated with more cognitive features of the tasks (e.g. Orata, 1928). A second, more telling, criticism of the identical elements theory concerned the assumption that stimulus–response pairs are the basis of transfer. This assumption was criticized as being so restrictive as to rule out transfer altogether (Meiklejohn, 1908). Unless a new task involved the same responses to the same stimuli as an original training task, there would be no transfer between the tasks. Thus learning to write with a pencil would be of no benefit for writing with a pen. Therefore, stimulus–response pairs were not sufficiently abstract to be identified as the identical elements of Thorndike's theory.

Little was achieved in the years that followed Thorndike's work to provide a suitable representation for skill. Research into transfer was more concerned with the effect of various training conditions on transfer performance. However, Briggs (1969) suggested that analysing the relationship between common features of tasks and the extent of transfer between them could only provide a preliminary understanding of transfer and that it was 'important to determine *what* is learned . . . during training . . . for a more complete understanding (p. 217)'. Hence, the progress of skills research was restricted by the lack of suitable concepts for describing the 'what' of learning.

The cognitive representation of knowledge became a popular topic of research and theory in the 1960s resulting in many suggested modes of representation, such as semantic networks (Collins and Quillian, 1969, 1972), productions (Newell and Simon, 1972), schemas (Minsky, 1975), and mental models (Johnson-Laird, 1983). As a result, theoretical tools not available to Thorndike were developed, enabling more abstract discussions of the processes underlying skills and transfer.

2.2.7 Fitts and the phases of skill acquisition

The above summary of results and theories was mainly concerned with motor skills. In fact, until the 1960s, skills research was almost exclusively concerned with motor tasks. The most influential definition of what constitutes skilled performance was restricted to muscular performance (Pear, 1948). It was not until the cognitive revolution during the 1960s that cognitive performance began to be considered under the skill heading. The most important early discussion of skill acquisition as involving cognitive processes was by Fitts (1964). This work is important in the sense that it specified the phenomena that a theory of skill acquisition should explain.

Although Fitts did not propose a theory of skill acquisition, his descriptions of the sequence of events involved in developing a skill were the first steps towards such a theory. One of his descriptions is worth quoting because it identifies the processes that need elucidation by theory:

> An adult, or even a child of a few years of age, never begins the acquisition of a new form of skilled behaviour except from the background of many already existing, highly developed, both general and specific skills. Thus, the initial state of our model is not that of a random model, but an already highly organised system processing language skills, concepts, and many efficient subroutines . . . . The number of such identifiable highly developed skills in an adult is certainly in the hundreds, each having its own executive program and library of subroutines, many of the subroutines being shared with other skills.
>
> The actual sequence of behaviour processes employed early in learning varies with the type of activity, of course, but might be somewhat as follows: The S observes or samples certain aspects of the environment, puts this information into short-term storage after some recoding, makes a decision such as selecting an appropriate subroutine which sets up a response pattern, executes a short behaviour sequence . . ., samples the internal and external feedback from this response plus additional stimulus information from the environment, recodes and stores this new information (in the process losing some of the information already in short-term storage), makes another decision which might be to use a different subroutine, and so on. As learning progresses, the subroutine becomes longer, the executive routine or overall strategy is perfected, the stimulus sampling becomes less frequent and the coding more efficient, and different aspects of the activity become integrated or coordinated . . . . As learning

> continues, overall performance may come to resemble more and more closely a continuous process. The overall program having now been perfected, frequent changes no longer need to be made in it. However, subroutines may continue slowly to become more efficient, and the S to become increasingly able to carry on the entire behaviour process while engaged simultaneously in other activities, with little or no interference between the two.
>
> (Fitts, 1964, p. 260)

Fitts (1964) suggested that skill acquisition involved three phases. The early phase was termed the 'cognitive stage' by Fitts. This stage lasts for only a few trials while the subject comes to terms with instructions and develops performance strategies. According to Fitts, these strategies develop from general 'sets' and strategies developed with previously learned tasks. Refinement of the performance strategy comes in the intermediate phase—'the associative stage'. Features of the previously learned strategies that are appropriate to the new situation are strengthened on the basis of feedback, whereas inappropriate features are weakened. This process forms new associations between specific stimulus cues and appropriate responses. In the end phase—'the autonomous stage'—the components of the performance strategy slowly become more autonomous so that they are less subject to cognitive control or external interference. As a result skilled performance of the task requires increasingly less processing, which means that more processing resources can be used for other activities. During this phase, skills continue to become faster and more efficient although the rate of improvement slows with practice.

Fitts provided no theoretical accounts of the processes he identified in the three phases, although he did point out where he thought existing theories were useful in this scheme. For instance, the selection of previously learned general sets and strategies for incorporation into new strategies draws on Crossman's (1959; see below) general probability learning model. In this model, subjects are presumed to possess a repertoire of methods for performing a task. Each method is selected at random and the probability of its subsequent selection is dependent on its performance speed—the faster ones being more likely to be selected. This process predicts the typical power function speed-up found in skill acquisition (e.g. Newell and Rosenbloom, 1981). However, other features of Fitts's description of skill development are not specified to the same extent. The development from the initial stages of task performance of the 'executive program', the 'subroutines', and the relationship between these is not given a clear process description. However, the suggestion that these skills have a hierarchical structure that is goal-driven is a large step towards such a description.

Fitts's work did not lead to a large increase of research into cognitive skills. Although there was certainly a growing interest in cognitive processes during this time, research was dominated by the advent of the information processing approach. Most researchers were interested in performance questions, where performance was examined in a limited context. As a result the field was soon well supplied with very specific theories concerned with isolated and disparate processes (e.g. Chase, 1973). There did not appear to be much interest in higher cognitive processes, that is, processes responsible for changes in performance.

Eventually, though, cognitive scientists began to be interested in more unified theories of cognition, attempting to describe cognitive architectures that could account for a wide variety of phenomena (e.g. Minsky, 1975; Newell and Simon, 1972). A number of these theories also consider the nature of skill acquisition, and these are discussed below.

2.3 **Theories**

In this section, we present the major theories of skill acquisition and evaluate the extent to which the theories meet the Skill Acquisition Phenomena Criterion. There are two issues that are of concern in this section. First, we focus on the various explanations for why practice leads to performance improvements, particularly improvements that have the negatively accelerated feature of power functions. Second, we look at how each theory handles the issue of transfer. A third issue—the effect of transfer on the shape of learning curves—is another feature of skill acquisition that the theories need to explain in order to meet the Skill Acquisition Phenomena Criterion. This issue is considered in Chapter 3.

A broad distinction can be made between theories of skill acquisition. On one hand, there are the theories that propose that practice leads to performance improvements because practice has the effect of refining procedures for performing a task (e.g. Anderson's ACT, Newell and Rosenbloom's SOAR). On the other hand, there are theories that view performance improvements as a by-product of some of the consequences of practice. For example, some theories claim that practice leads to greater knowledge (i.e. more memories to call upon) which in turn leads to faster performance (e.g. Logan's Instance theory). This distinction, is not exhaustive, as recent models of skill acquisition have begun to blur the boundaries between strategy refinement and memory-based performance as explanations of performance improvement (e.g. Rickard, 1997, 1999; Wenger, 1999). In this section, our presentation of the theories maintains the distinction; however, throughout the book there will be cause to consider abandoning such a clear separation of views.

The aim of this section is to begin to evaluate the theories in terms of whether they satisfy the Skill Acquisition Phenomena Criterion for suitability as theories of skill acquisition. The main issues in focus in this section are explanations provided by each theory with respect to why learning curves are typically negatively accelerated, and what features determine transfer.

2.3.1 Skill acquisition as strategy refinement

Crossman

Crossman (1959) provided an early view on how practice leads to performance improvements. He suggested that, when faced with a new task, we have many strategies that can potentially be used. With practice at the task, we monitor the outcomes we produce as a result of the various strategies used. With time we come to favour the more efficient strategies. These are used more often and this produces performance speed-up. Thus, Crossman's model qualifies as an example of a theory that proposes that practice leads to more efficient procedures for performing a task. The theory does not suggest that practice modifies strategies to make them more efficient. Instead, practice leads to the selection of the most efficient strategy among several.

Crossman's theory provides an account of the power law of learning. According to the theory, it is easier to find faster, more efficient methods early in practice, and so large gains in performance time are more likely at this stage. As practice continues, more efficient methods become harder to find, and so performance time improves by ever-smaller amounts.

Crossman's theory is not comprehensive on the topic of transfer. Obviously transfer is important to the theory as it relies on a range of strategies being available for performance of new tasks, however, Crossman did not deal with the issue in any great detail beyond a re-statement of Thorndike's identical elements hypothesis: '. . . transfer of skill from one task to another will take place where methods appropriate to one are also appropriate to the other' (Crossman, 1959, p. 163).

Crossman's model did not specify the origin of the various strategies. Furthermore, he did not provide evidence to support the claim that people will tend to follow the more efficient strategy—certainly there is evidence that this does not occur (e.g. Luchins, 1942). Thus, Crossman's theory cannot be considered as a serious contender for a comprehensive theory of skill acquisition. Nonetheless, taken with the other theories presented below, it illustrates that there are many ways in which particular behaviour changes (i.e. performance improvements with practice) can be modelled.

ACT-R

One of the most influential theories of skill acquisition, and to date, the most comprehensive, is the ACT theory, in its various incarnations. Early versions of this theory include ACTE (Anderson, 1976)—a successor to the HAM model of memory (Anderson and Bower, 1973), ACTF (Anderson *et al.*, 1979, 1980), and ACT* (Anderson, 1982, 1983a). ACT-R (Anderson, 1993; Anderson and Lebiere, 1998) is the latest version of the theory and inherits most of the features of the earlier versions. ACT-R is first and foremost a general theory of cognition that describes an architecture that underlies all cognitive processes. The ACT-R theory, therefore, qualifies as a candidate for a unified theory of cognition, as advocated by Newell (e.g. 1989, pp. 404–405). It is from this general theory of cognition that the ACT-R theory of skill acquisition emerges. Features of the cognitive architecture proposed by ACT-R that are relevant to skill acquisition are described below. This is followed by a description of the processes involved in skill acquisition and transfer.

The cognitive architecture proposed by ACT-R ACT-R proposes that the architecture underlying cognition is a production system. Such a system involves the application of production rules that are activated by the contents of a working memory. In this sense production systems can be considered 'cognitive S–R (stimulus–response) theories' (Anderson, 1983a, p. 6). Production systems are not peculiar to the ACT-R theory and have the property of being computationally universal. That is, they can be used to model all cognitive processes. However, the use of a production system in ACT-R is not simply to re-describe behaviour but is sensitive to psychological and empirical constraints. As a result, testable predictions can be made on the basis of this production system.

Basic to ACT-R is the distinction between declarative and procedural knowledge. Declarative knowledge can be considered to be the representation of facts (e.g. 'A red traffic light is a signal to stop'.). Procedural knowledge is basically the representation of what to do in particular situations. Thus, actions are contingent on certain conditions being present. In ACT-R, procedural knowledge is embodied as production rules—condition–action pairs that associate the presence of a particular data pattern in working memory (condition) with the performance of a certain action. Therefore, when the condition of a production rule is satisfied, the production can apply and the action will follow (e.g. 'IF traffic light is red, THEN stop').

Particular features of the ACT-R production system constrain the activation of productions. These are as follows: (1) Productions can only be activated by knowledge that is currently active in working memory. (2) The speed at which the condition of a production is matched to data in working memory is a function of the strength of the production. In ACT-R, productions gain strength with successful application.

In addition to the above constraints on production activation, ACT-R includes several rules of conflict resolution. Originally, in ACT*, there were three rules that determined which production will apply when the conditions of more than one production are matched by the data in working memory. (1) Refractoriness prevents the same production from applying to the same data in the same way more than once. As a result, the production cannot repeat itself over again. (2) When two or more productions can apply, the production with the more specific condition will apply. (3) Specificity and strength interact in the application of productions. If a production with a general condition is selected and applies before a more specific production can apply, then the general production will be the one that actually applies. Therefore, more specific productions can only take precedence if they have sufficient strength to ensure faster selection and application times than more general productions. This strength comes only with successful application (i.e. practice). ACT-R now includes one overriding conflict resolution mechanism. When several productions share the same goal, the most useful production is chosen for execution. Usefulness, or the expected gain of a production, is defined as:

$$\text{Expected Gain (E)} = PG - C$$

where *P* represents the probability that a production will satisfy the goal if it is executed, and *G* is the worth of the goal. Thus *PG* represents the potential gain of firing the production. *C* represents the cost of executing the production, and this generally reflects the time to achieve the goal. Hence E reflects the expected net utility of executing a particular production. When several productions compete for the same goal, the one with the highest expected gain will fire.

Individual productions do not usually function in a vacuum. Instead, sets of productions are organized with hierarchical goal-structures. This organization accounts for the hierarchical nature of human behaviour identified by Fitts (1964). Productions that underlie the performance of a particular behaviour are organized around the satisfaction of goals and sub-goals. This results in serial processing where only one goal can be attended to at a time (Anderson, 1983a, p. 33). This goal-driven structure also has the function of biasing pattern-matching processes towards matching structures involving the current goal. Considering the specificity constraint described above, this means that productions that refer to the current goal are more likely to apply and apply more rapidly than productions that do not refer to this goal.

The above description of the production system underlying the ACT-R theory of cognition is only a brief one, touching upon those features that are important for understanding the ACT-R theory of skill acquisition. A more complete account is available in Anderson and Lebiere (1998).

The ACT-R theory of skill acquisition The ACT-R theory of skill acquisition was developed by Anderson and his colleagues after studying the learning of geometry proofs (Anderson, 1982, 1983a, 1989a; Anderson *et al.*, 1981, 1993a; Neves and Anderson, 1981), computer programming in the LISP language (Anderson, 1986, 1987, 1989a; Anderson and Reiser, 1985; Anderson *et al.*, 1985, 1989, 1993b; McKendree and Anderson, 1987; Pirolli and Anderson, 1985a; Singley and Anderson, 1989) and other computer languages (Anderson, *et al.*, 1993c), text-editing (Anderson, 1987; Singley and Anderson, 1985, 1989), calculus (Singley and Anderson, 1989), algebra (Anderson, 1989a; Blessing and Anderson, 1996), language processing (Anderson, 1982, 1983a), schemata (Anderson *et al.*, 1979), problem-solving (Anderson and Fincham, 1994; Anderson *et al.*, 1997, 1999), and creating computer simulations of the processes underlying skill development in each of these areas. The computer simulations relied on similar architectures to that specified by ACT-R (and previously ACT*) and so provided a useful method of assessing the adequacy of the ACT-R account of skill acquisition.

In all of the work cited above involving Anderson, the basic theory of skill acquisition has remained consistent, varying little from the first appearance of the ACT* theory. However, some details have been modified recently and these are described below. The general sequence of events in skill acquisition is suggested to be as follows: Knowledge relevant to the performance of a skill begins in declarative form. This knowledge is interpreted by general productions called weak problem-solving methods. These methods are termed 'weak' because they are domain-general, that is, their operation is not specific to any particular type of task (e.g. analogy). Domain-specific productions are created by a process called compilation. This process involves two sub-processes. The first is proceduralization, which describes the creation of domain-specific productions as a by-product of the interpretation of declarative knowledge via weak problem-solving methods. These new productions perform the goal behaviour without the need to consult declarative knowledge. Composition is the second compilation process, and describes the formation of efficient productions by collapsing sequences of productions into single productions that have the effect of the series. The likelihood of a production being applied in a particular situation, and the speed at which the production will be executed, are both functions of the production's strength. Productions accumulate strength depending on their history of success. Stronger productions are matched and applied faster. Therefore, highly practised productions are executed faster than newly formed productions. All of the above processes will now be described in more detail.

The processes underlying skill acquisition will be illustrated with respect to a student learning to solve algebra problems. The strategy described is not necessarily how people solve such problems but is useful in illustrating the changes that can occur with practice.

Imagine that a teacher is describing an algebra solution method to a student. The teacher may start with a problem like $79 = 3x + 4$ and tells the student that the goal is to solve for 'x'. To achieve this goal requires achieving a number of sub-goals. For example, the teacher may tell the student that the first step in realizing the overall goal is to isolate the 'x' term on the right-hand side of the equation. This will mean eliminating the '4' from this side of the equation. The teacher will then demonstrate how this is done, by adding '-4' to both sides of the equation:

$$79 + (-4) = 3x + 4 + (-4)$$
$$\Rightarrow 75 = 3x$$

Having achieved this sub-goal of isolating the 'x' term on the right-hand side of the equation, the teacher may then describe the second sub-goal, which is to eliminate the coefficient of the 'x' term, which is 3. This is achieved by dividing both sides of the equation by 3:

$$\frac{75}{3} = \frac{3x}{3}$$
$$\Rightarrow 75 = 3x$$

This is then the solution to the problem.

The student's memory of these instructions for how to solve the problem can be considered declarative knowledge. It represents knowledge about how to solve that particular problem, but cannot be used alone to solve other problems. For this knowledge to be useful in solving other problems requires processes that can interpret this knowledge and translate it into action. In ACT*, these processes were described as weak problem solving methods that can be useful in a wide range of domains. According to Anderson (e.g. Singley and Anderson, 1989), humans develop these at an early age and by adulthood these methods are well developed. These problem-solving methods include analogy, means–end analysis, hill climbing, and pure forward search.

In the algebra example, if the student were presented with another problem to solve, such as $85 = 4x + 5$ analogy would be the most likely method to apply. Analogy would function to enable the student to mimic the previous solution. This process will only apply, of course, if the student notices the usefulness of the previous solution (e.g. Gick and Holyoak, 1983;

Holyoak, 1985). The process can be illustrated with the following plausible imitation of a talk-aloud protocol (for real examples of such protocols, see Anderson, 1983a).

'This problem ($85 = 4x + 5$) looks similar to the teacher's example, so maybe if I try the method that the teacher described I'll solve the problem. The teacher started by isolating the 'x' term on the right-hand side of the equation by adding the negative of the leftover term (4) to both sides of the equation. In this new problem the leftover term is 5, so I should add -5 to both sides of the equation:

$$85 + (-5) = 4x + 5 + (-5)$$
$$\Rightarrow 80 = 4x$$

'After doing this, the teacher eliminated the number in front of the 'x' term by dividing both sides of the equation by that number (3). In this new problem the number in front of the 'x' term is 4 so I should divide both sides of the equation by 4:

$$\frac{80}{4} = \frac{4x}{4}$$
$$\Rightarrow 20 = x$$

So, the solution must be 20'.

Although this is a purely fictitious protocol, it captures the essence of protocols reported by Anderson (1983a). This example was designed to illustrate how general interpretive methods can be used to translate declarative knowledge into action, given the limitations of declarative memory (i.e. that the student can remember all that the teacher did and why) and that a previous solution will be noticed as useful (see Singley and Anderson, 1989, p. 34).

A by-product of the application of weak problem-solving methods to interpret declarative knowledge and achieve a solution is the formation of new productions. In contrast to the domain-general weak methods, these new productions are domain-specific. That is, their application is peculiar to the domain of the problem, operating only on particular features of that domain. This development of productions from the application of general methods is called compilation and, as indicated earlier, involves two processes: proceduralization and composition.

Proceduralization eliminates the reference to declarative knowledge by building into productions the effect of that reference. In the algebra example, proceduralization of the problem solution would mean that the student no longer needs to refer to the memory of the teacher's instructions to solve further problems. Instead, the student will have developed a set of productions that will solve such problems directly. These productions are domain-specific—they operate only in the domain of solving such algebra problems. This contrasts with the weak methods that will apply in a large variety of domains. An example of a set of domain-specific productions for solving the algebra problems is presented in Table 2.1.

The development of productions, such as those in Table 2.1 precludes the need to hold declarative information (i.e. the teacher's instructions) in working memory and use analogy at each step of the problem. Thus, to solve new algebra problems, the student does not need to continually refer back to previous solutions for directions. This prediction is supported by the dropout of verbal rehearsal of problem-solving steps that characterizes early performance on this type of task (e.g. Anderson, 1983a). As the need to refer to declarative knowledge is reduced with proceduralization, so should the load on working memory be similarly reduced. This prediction is also consistent with observation (e.g. Woltz, 1988).

An important feature to note of the set of productions in Table 2.1 is that they have a hierarchical structure that matches the goal structure implicit in the solution of such problems. This is fundamental to the ACT-R description of skills and skill acquisition and underlies the form of the production sets that develop. The second process involved in compilation—composition—is determined by the goal structure of problems. Composition collapses several productions into a single production. These productions must occur in a sequence and share the same overall goal. The new single production does the

Table 2.1 Example of a set of domain-specific productions for solving algebra problems

P1	IF	goal is to solve for x in equation of form a = bx + c
	THEN	set as sub-goal to isolate x on RHS of equation
P2	IF	goal is to isolate x on RHS of equation
	THEN	set as sub-goals to eliminate c from RHS of equation and then to eliminate b from RHS of equation
P3	IF	goal is to eliminate c from RHS of equation
	THEN	add −c to both sides of equation
P4	IF	goal is to eliminate b from RHS of equation
	THEN	divide both sides of equation by b
P5	IF	goal is to solve for x in equation and x has been isolated on RHS of equation
	THEN	LHS of equation is solution for x

work of the sequence but in fewer steps. For example, productions P2, P3, and P4 in Table 2.1 could be composed to:

```
P6    IF       goal is to isolate x on RHS of equation
      THEN     add −c to both sides of equation
               and then divide both sides of equation by b
```

With further practice, productions P1, P6, and P5 could be composed to:

```
P7    IF       goal is to solve for x in equation of form
               a = bx + c
      THEN     add  −c to both sides of equation
               then divide both sides of equation by b
               and result is solution
```

An algebra ‘expert’ (i.e. someone with many years of experience solving such problems) should be able to recognize this solution immediately upon observation of the problem. The expert would be unlikely to consider the intermediate steps that the novice needs to perform.

Compilation predicts a speed-up in performance for a number of reasons. First, Anderson (1982) suggests that the time to perform a task is a function of the number of steps involved. Therefore, since composition reduces the number of steps (productions) required to perform a task, with practice performance will rely on fewer productions and so will take less time. A more significant reduction in performance time comes with proceduralizsation. Performing a task on the basis of a set of productions that execute the task directly should take less time than having to interpret declarative knowledge for procedural directions. This accounts for the dramatic improvements in performance time observed by Singley and Anderson (1989) after only one trial of learning. Singley and Anderson reported that subjects showed a 50% improvement in the time to produce a LISP function from the first trial to the second trial.

The combined speed-up in performance predicted by composition and proceduralization is not sufficient to account for all the speed-up that is observed in skill acquisition (Anderson, 1982). ACT-R includes a tuning mechanism—strengthening—that results in further improvements in performance time with practice. Productions are strengthened with successful practice. That is, each time a production is applied successfully it gains strength. Conversely, if a production is applied inappropriately it loses strength. The stronger a production is, the faster it is to apply. So, the combination of compilation and strengthening predicts a speed-up in performance that continues with practice.

Anderson (1982) demonstrates how the compilation and strengthening mechanisms account for the power law of learning (e.g. Newell and Rosenbloom, 1981). The basic ACT* version of such a power function is:

$$T = N P^c$$

where N is the time on trial 1, related to original number of productions, P is the amount of practice, c is the rate of learning, $c < 0$; $c = f + g$ where f is the fraction by which the # of steps is reduced by composition, g is related to memory decay.

The ACT-R account of the power law of learning is examined in more detail in Chapter 3 when we consider the effects of transfer on the shape of learning curves.

Early versions of the ACT* theory of skill acquisition (e.g. Anderson, 1982, 1983a) included additional tuning mechanisms to strengthening—generalization and discrimination. These mechanisms were suggested to be automatic induction processes that refined the productions developed by compilation. Generalization is a process whereby general productions are generated from more specific ones. An example from Anderson (1987, p. 205) illustrates this process with respect to language acquisition. If a child has developed the following two productions:

IF	the goal is to generate the present tense of KICK
THEN	say KICK + s
IF	the goal is to generate the present tense of HUG
THEN	say HUG + s

the generalization mechanism would develop a more general rule that would be applicable in the above situations and others:

IF	the goal is to generate the present tense of 'verb'
THEN	say 'verb' + s

Discrimination has the effect of restricting the range of such general rules. Thus, in the above example, the general rule is overly general, and in certain circumstances is not appropriate. As a result, the discrimination mechanism would generate new rules appropriate for these circumstances. For example,

IF	the goal is to generate the present tense of 'verb' and the subject of the sentence is singular
THEN	say 'verb' + s
IF	the goal is to generate the present tense of 'verb' and the subject of the sentence is plural
THEN	say 'verb'

Despite the utility of generalization and discrimination in accounting for various phenomena in language acquisition (e.g. Anderson, 1983a), later versions of ACT* (Anderson, 1986, 1987, 1989a; Singley and Anderson, 1989) suggested that these two tuning mechanisms are unnecessary in a theory of skill acquisition. In fact, the effects of these mechanisms can be implemented by the same general problem-solving methods that initiate proceduralization (e.g. analogy) (Anderson, 1987). In addition, by having generalization and discrimination implemented by these general methods, the induction of more refined rules is more sensitive to semantic and strategic factors than the automatic processes of the generalization and discrimination mechanisms (Anderson, 1987).

A further refinement of the ACT theories has seen the composition process of ACT* dropped. This is not to say that composition-like phenomena do not occur. Indeed, there is behavioural evidence to suggest that people can acquire multistep productions, such as P7 above (e.g. McKendree and Anderson, 1987). Anderson (1993) claims, that there is no evidence that this form of production refinement occurs automatically with practice. Instead, in ACT-R, analogy to examples that illustrate a more direct solution strategy can lead to new, similarly direct, productions. These new productions, will compete with old productions that may perform the task in more steps, but are faster because of prior practice. Hence, the relative expected gain of the old productions and the new, more refined productions will determine which productions are executed. Thus, practice will not automatically lead to composed productions.

The removal of an automatic composition mechanism in the ACT theories means that performance improvements that result from practice are mainly due to strengthening of productions. In ACT* practice also led, automatically, to greater refinement of productions (e.g. composition). In ACT-R this refinement may occur, but will be determined by considerations of the gain and cost of performance. The only automatic consequence of practice in ACT-R is that productions are strengthened. Because the strength of productions increases as a power function of the number of executions, and productions are executed faster as their strength increases, practice leads to faster performance in a manner described by the power law of learning. In other words, skill acquisition is a process whereby practice leads to the development of procedural knowledge (productions) that is altered by repeated execution (i.e. it is strengthened), and this alteration in turn leads to faster performance. Thus, it is a change in the characteristics of knowledge that results in better performance.

In addition to the observations of cognitive skill acquisition that were mentioned above, the ACT-R theory of skill acquisition provides a useful account of the three phases of skill development described by Fitts (1964).

ACT-Rand Fitts's phases of skill acquisition The early phase of skill acquisition identified by Fitts (1964) corresponds in ACT-R to the application of general problem-solving methods to declarative knowledge and to the initial development of productions. Fitts suggested that this phase only lasts for a short time, which is consistent with Anderson's reports of one-trial learning (e.g. Singley and Anderson, 1989). Fitts describes this phase as the cognitive stage, where most of the thinking about a task is performed. Anderson (1983a) claims that it is natural to equate this stage with the interpretive application of knowledge. Certainly there is considerable evidence that higher order cognitive activities, such as comprehension of task requirements and planning, are more prevalent early in skill development than later (e.g. Ackerman, 1988; Woltz, 1988). In addition, processing during this stage is more error-prone and deliberate than in subsequent stages (Ackerman, 1988; Woltz, 1988) as working memory resources are stretched by the interpretation of declarative knowledge.

The intermediate phase identified by Fitts describes the formation of specific associations between stimulus cues and appropriate responses. The similarities between such associations and production rules are obvious. In ACT-R, this stage corresponds to the dropout of verbal rehearsal of instructions and the associated reduction of working memory load.

The end phase was described by Fitts and Posner (1967) as the stage where 'component processes become increasingly autonomous' (p. 14). During this stage, skills are less reliant on working memory resources and become faster with practice. The ACT-R theory suggests that productions gain strength with practice and this results in faster application of the productions. In accordance with the power law of learning, the effect of strengthening on improvement becomes increasingly small with practice. Eventually, after many thousands of trials (Anderson, 1989b), no further improvement will be observed. During this phase, performance may appear to be automatic (e.g. Shiffrin and Schneider, 1977): when the conditions of a production are satisfied, the action will follow automatically.

Transfer The ACT-R theory of skill acquisition also describes how transfer can occur between tasks. Anderson (1987; Anderson and Singley, 1993; Singley and Anderson, 1989) resurrected the identical elements theory by identifying production rules as the elements of knowledge underlying transfer. According to this proposal, transfer between two tasks is determined by the extent to which productions underlying performance in one task are useful in performing the second. The greater the production overlap, the greater the transfer.

Productions have an immediate advantage over stimulus–response pairs with respect to representing the processing operations underlying skilled performance: productions are abstract cognitive representations. As described above, productions are formed as the by-product of an interpretive process that compares two declarative representations. For example, analogy compares the representation of a previous solution to the representation of the current situation and extracts common features. The productions that result from this process are generalizations and, therefore, are necessarily abstract (Singley and Anderson, 1989).

Three forms of transfer are traditionally considered in any discussion of transfer results: positive, negative, and zero transfer. The identification of identical productions as the basis for transfer leads to strong predictions concerning each of these situations. These are described below, along with studies that examine these predictions.

Positive transfer As a task is practised, a set of productions is developed that underlies the performance of this task. If the performance of a second task can utilize these same productions then positive transfer will result between the two tasks. Positive transfer in this situation refers to the fact that knowledge developed in one situation is transferred to another situation. This has a positive effect in the sense that there is no need to develop productions from scratch to perform operations that existing productions already perform. Therefore, the extent of this positive transfer is determined by the number of productions developed in the context of the first task that can be used in performance of the second task.

Singley and Anderson (1985, 1989; Anderson and Singley, 1993) provided support for the ACT-R account of transfer in a detailed study of text-editing skill. Two basic types of editors were examined: line editors that allowed examination of only one line of text at a time, and screen editors that enabled viewing of whole screens of text. Singley and Anderson predicted that there would be almost complete positive transfer between two line editors and only partial transfer from line editors to screen editors. The transfer predictions were based on models of the productions underlying performance with each of the editors. Despite the fact that the two line editors shared few commands, Singley and Anderson identified considerable production overlap in these editors, mainly concerned with the more abstract planning operations. This was shown to result in almost complete transfer between the two line editors: two days of practice with one of these editors was almost equivalent to two days of practice with the other in terms of preparation for further performance with this second editor. Much less production overlap was identified between the line editors and the screen editor and this was shown to result in only partial transfer from training with the line editors to performance with the screen editor.

Kieras and Bovair (1986) have also reported success at predicting the extent of transfer between tasks on the basis of the number of shared productions. Detailed predictions of the training time to reach a performance criterion were shown to be accurate when the task was analysed in terms of old and new productions. Old productions had been learned in the context of another task and so did not require learning. New productions were those that needed to be developed. Kieras and Bovair predicted and observed that the greater the number of old productions involved in the performance of a new task, the smaller the training time required on this task.

Negative transfer Negative transfer refers to a situation where performance of a task is worse than if a preceding task had never been performed. One interpretation of such a result is that the procedure for performing the second task is performed less effectively as a result of another procedure having been learned in the context of the first task (Anderson, 1987). However, the ACT-R account of transfer predicts that negative transfer in this sense does not exist (Singley and Anderson, 1989). The worst case that can be expected in a transfer situation is that two tasks share no productions, resulting in zero transfer (this situation is described below). ACT-R proposes that results that might indicate negative transfer in fact indicate positive transfer. However, what have been transferred are non-optimal methods for performing the task. Therefore, productions developed in the context of one task are transferred to the performance of another task. These productions do not interfere with the execution of productions relating to the second task (i.e. negative transfer). Instead, when these productions are used to perform the second task, performance is less efficient than if it relied on productions that were developed in the context of the second task only. This situation will occur when the stimulus conditions of the second task match the conditions of productions developed in the context of the first task. These productions will execute in response to these stimulus conditions, although the processing strategy they embody will not be optimal for the second task.

Some evidence consistent with the ACT-R account of negative transfer exists, the most famous being the demonstration of the *Einstellung* phenomenon by Luchins (1942). Subjects were required to solve a number of problems that involved determining the best method of combining three water jugs (A, B, and C) to make up designated volumes of water. The problem sequence was such that subjects were presented with five problems that each had the same solution (i.e. combine the jugs according to B - A - 2C). Following these problems, the subjects were presented with another two problems that could either be solved with this same method or with a more direct method (A + C).

Luchins found that almost all of the subjects solved the two transfer problems with the method discovered during training. In contrast, control subjects who were first presented with the transfer problems almost always solved them with the easier method. Therefore, the training problems encouraged the use of a non-optimal solution procedure with the transfer problems. The stimulus conditions of the transfer problems were obviously sufficiently similar to those of the training problems to trigger the execution of this non-optimal method. This meant that subjects could utilize a previous solution rather than develop a new strategy specific to the transfer problems. Further evidence that negative transfer is in fact the positive transfer of non-optimal methods has been reported with text-editing (Singley and Anderson, 1989) and computer programming (Kessler and Anderson, 1986).

Zero transfer The ACT-R account of transfer predicts that training with one task will provide no benefit for performance with a second task if the two tasks do not share common productions. More specifically, transfer between tasks is restricted to common productions and is not related to the abstract knowledge underlying the productions. Therefore, there will be no transfer between skills that use the same knowledge in different ways. This proposal leads to some counter-intuitive predictions. For example, there should be no transfer between the comprehension and generation of language. Singley and Anderson (1989) reviewed the small amount of research relevant to this issue and concluded that there was some evidence that comprehension and generation involved separate systems. In addition, Singley and Anderson reported an experiment that examined the training of subjects to solve calculus problems. Results of this experiment suggested that translating written problems into equations was completely independent of the solution of these equations. These processes shared common abstract knowledge about the function of calculus operators but used it in different ways. Similar evidence was provided by McKendree and Anderson (1987), who found that there was little transfer between evaluating and generating LISP code. There also appeared to be little transfer between evaluating certain combinations of LISP commands when the same components were presented in one combination during training and another during transfer. Thus although the same primitive knowledge was used in both cases, apparently the productions developed in one situation were not appropriate for the other situation. These studies demonstrate that the acquisition of skills is specific to their use.

In conclusion, by specifying productions as the elements underlying transfer, the ACT-R account of transfer provides an advance over Thorndike's identical elements theory. Productions appear to be more appropriate representations of procedural knowledge than stimulus–response pairs. In addition, there is some evidence that transfer between tasks can be predicted on the basis of identical productions underlying performance of the tasks.

Soar

In several respects the SOAR theory (Laird, *et.al*. 1987; Newell, 1990; Rosenbloom *et al.*, 1991) is similar to the ACT theory. It is designed to provide accounts of a wide range of cognitive phenomena, and indeed has been successful in a number of domains (e.g. simple reaction tasks, the Sternberg item recognition paradigm, typing, verbal learning, cryptarithmetic, syllogisms, and sentence verification; Newell, 1992). In fact, SOAR has been proposed as a candidate for a unified theory of cognition (Newell, 1990). SOAR has been implemented as a computer program, enabling researchers to simulate phenomena and so compare observed and predicted data. Finally, SOAR is instantiated as a production system. That is, SOAR holds that behaviour results from the execution of a series of production rules.

In SOAR, all tasks are problems. Problems are solved by the execution of productions. Productions are executed if their conditions match task conditions or the actions of satisfied productions. There is no explicit conflict resolution mechanism in SOAR, although a pseudo-conflict resolution procedure does operate. This procedure occurs as part of the decision cycle. In each decision cycle, all productions that can be executed do so. However, if several productions fire in response to the same conditions, they do not all control behaviour. The execution of a production involves adding new knowledge to long-term memory. This knowledge will include preferences about which decisions or behaviours are better than others. The end of each decision cycle involves sorting through the preferences to determine the most appropriate behaviour. Performing this behaviour then triggers the next decision cycle.

Often the productions in long-term memory will not be sufficient for determining which step to take next. In this case an impasse occurs. SOAR treats such impasses as just another problem and sets subgoals to resolve the impasse. SOAR possesses several strategies for dealing with impasses, but each is designed to acquire knowledge about the appropriate next step. Impasses can occur while the system is already working on a sub-goal. As a result, procedures for performing complex tasks will typically possess a hierarchy of goals and sub-goals.

Encountering and resolving impasses provides the opportunity for learning. According to SOAR, when an impasse is resolved, a new production is created. The condition of this production represents the relevant working memory elements at the time of the impasse, and the action represents the solution taken. Thus, if the same situation is encountered again, it is not necessary to go through the problem-solving process again to resolve the impasse. Instead, SOAR suggests we simply retrieve and execute the previously successful solution.

Remembering past solutions to problems is known as chunking in SOAR. Being able to recognize a problem situation and execute a direct solution has obvious performance benefits compared to having to solve the same problem again each time it is encountered. Furthermore, in complex tasks, chunking can occur many times, effectively chunking together the behaviour of several chunks into new chunks. Thus, each time the same task is performed, there is less problem solving and more direct retrieval of solutions. In this way, SOAR can provide an account of the power law of learning. In essence, the negatively accelerated feature of the power law comes about because, in the initial stages of practice, there are many opportunities for chunking to occur, and so large gains in performance speed can result. As practice continues with a task, there is less opportunity for chunking to occur, and so less potential for further improvement in performance time. It is important to note that SOAR accounts for the negatively accelerated feature of learning curves with its chunking mechanism only and, unlike ACT, does not invoke a strengthening mechanism.

The SOAR account of transfer is similar to the ACT account. The productions that result from chunking are abstractions of the problem-solving process, and so may apply beyond the conditions under which they were originally acquired. However, productions can only be executed when their conditions are satisfied. Thus, SOAR resembles ACT in following an identical elements view of transfer: Transfer between two tasks will occur to the extent that productions developed to perform one task can be utilized in performance of the other task.

MacKay

Originally designed to provide an account of possibly the most complex human skill, speech production, MacKay's (1982) theory is also applicable to the full range of skilled behaviour. MacKay explained how a cognitive event (e.g. an idea to be expressed) could be translated into a muscular response (e.g. an utterance reflecting the idea) by proposing that such skilled behaviour is represented as a network of interconnected nodes. The network is hierarchical in nature, such that if an idea is to be expressed in speech, it is originally conceived in a conceptual system as a series of propositions. These propositions are then connected to a phonological system where they are translated into phonological information. In turn, the phonological information is translated into information about which muscles to move to make the sounds in a muscle movement system.

Greater fluency in behaviour comes about with practice, which results in connections between nodes being strengthened. Thus repeating a sentence aloud several times results in stronger connections between the propositional nodes, the phonological nodes, and the muscular nodes that represent the various features of the sentence. In essence, stronger connections between nodes means that activated nodes can prime those nodes to which they are connected lower in the hierarchy. This priming results in faster activation of the lower nodes, and so leads to greater fluency.

MacKay's (1982) theory accounts for the power law of learning in the following manner: The rate at which a node can be activated is related to the rate at which it can be primed. Priming rate is an exponential function of the frequency with which the node has been activated by a connected node in the past. Thus, an unpracticed node will show rapid improvement with practice, nodes with intermediate practice will show moderate improvement with further practice, and those nodes with extensive priming histories will be unlikely to have room for further improvements. The speed with which a task can be performed is a function of the performance of all of the nodes involved in performing that task. Thus, early in practice with a task, performance is dominated by nodes with little prior practice, and hence performance improvements will be large. As practice proceeds, learning rate becomes slower because all nodes are further along their learning curves and so are only improving slowly.

MacKay's (1982) theory also provides an account of transfer of training. To some degree the theory subscribes to the identical-elements view. In this theory, the elements are the nodes that represent particular aspects of a skill. Thus, the extent to which practice on one task will transfer to performance of a second task is determined by the extent to which both tasks engage the same nodes. For example, consider a bilingual person who repeats a sentence in one of their fluent languages several times. Then they are asked to read aloud a translation of that sentence in their other language. MacKay reports that there is considerable transfer in this situation, such that the translation sentence is read faster than control sentences. MacKay attributes this result to the fact that the original and translation sentences share nodes at the conceptual level (i.e. they both express the same idea). Although they may not share many nodes at the phonological and muscular level, nodes at these levels are so over-practiced that further practice is unlikely to result in performance improvements that would affect transfer.

Connectionist models

The final set of models to be considered in this section arguably does not belong in this section for several reasons. However, the models do possess some features in common with the other models in this section, and virtually no features in common with models in the next section. For this reason, connectionist models are discussed here and the appropriateness of including them in the category of 'strategy refinement' models is considered.

Connectionist models are models of memory processes that are based on some features of neural tissue (e.g. McClelland and Rumelhart, 1986; Rumelhart and McClelland, 1986). That is, these models involve networks of interconnected units, with each unit being activated by the firing patterns in units to which it is connected. Once a unit is activated, it can pass on this activation to other units. Typically, connectionist models have several layers of units. The input layer will be designed to represent features of the environment. The output layer will be designed to pass on features of the memory system to the environment. Some models also include one or more hidden layers between the input and output layers. Knowledge in the system is contained in the connection strengths between the units. There are three types of connection strength. A connection strength of zero means that activation of the connection has no effect on the units at the end of the connection. A connection strength with a positive weighting means that input has an excitatory effect, whereas a negative weighting means that input has an inhibitory effect. When input is received by the system, the various connection strengths result in firing patterns among the layers within the network. These firing patterns represent the system's response to particular environmental stimuli. Depending on the situation this response could be the perception of a word, the detection of an instance of a category, or a motor response to be executed.

Research involving connectionist models is often concerned with training a system to respond in particular ways to particular stimuli. In other words, the connectionist models can be configured to learn about an environment. Many different learning algorithms have been developed for use in connectionist models, but the majority appear to include some mechanism whereby connection strengths are modified to reduce error between the system's output and the desired output. Thus, the fact that connectionist models can simulate learning makes them relevant to the current discussion of skill acquisition. Furthermore, because learning in these models involves changes to existing representations (i.e. connection strengths), this learning can be considered to involve a form of strategy refinement in that one form of response is modified by experience to closer approximate the optimal form of response.

Having made the case for inclusion of connectionist models at this point in the discussion, it turns out that very few models have been proposed as accounts of the development of skilled behaviour. If we follow the criteria used so far, there are, to our knowledge, no connectionist models designed to account for both negatively accelerated improvements in performance time with practice, and transfer. There are some models, however, that deal with the first issue.

One connectionist model of skill learning has been proposed by Cohen *et al.* (1990). This model was specifically designed to provide an account of the classic Stroop effect (Stroop, 1935). In the Stroop task, words are presented in various ink colours. People are asked to state the colour of the ink as quickly as they can. The Stroop effect is observed in conditions where colour names are presented in other ink colours (e.g. RED presented in blue ink). Naming times in such conditions are typically slower than in conditions where words other than colour names are presented. The usual explanation for this effect is that reading words is an automatic behaviour, a skill resulting from many years of practice. When faced with a colour name, people have difficulty suppressing the typical response (i.e. 'red' for RED) in order to provide the correct response (i.e. 'blue' when RED is presented in blue ink). The Cohen *et al.* model was designed to simulate the development of an automatic skill (i.e. reading) that would interfere with the execution of another response (i.e. colour naming). Using the activation of units corresponding to particular responses to represent reaction time (i.e. the greater the activation, the shorter the reaction time), the model was successful in simulating the Stroop effect.

In addition to simulating the Stroop effect, the Cohen *et al.* (1990) model exhibited power function learning. The model was able to produce this result by virtue of the way in which connection strengths were modified during learning. Early in practice, there is likely to be a large difference between the actual output of the system and the desired output. As a result, large changes are made to the connection strengths. As practice increases and the system learns, the size of the discrepancy between the system output and the desired output will diminish. Connection strengths are then altered accordingly by ever decreasing amounts. This pattern of change in connection strength, together with the model directly relating reaction time to connection strengths (i.e. the stronger the connection, the shorter the reaction time), produces the negatively accelerated reduction in reaction times characteristic of the power law.

No connectionist model of skill learning currently provides an explicit account of transfer. Cohen *et al.* (1990), for example, did not consider the issue of transfer, either between different stimuli or different tasks, so the ability of their model to handle such situations is unknown. Connectionist models in general, appear well suited to dealing with such situations. Certainly, models do exist that provide an account of the ways in which behaviour is affected by changes in stimuli between training and transfer (e.g. Ratcliff, 1989; Willshaw, 1989). It may simply be a matter of time before a connectionist model is proposed that provides an account of both skill acquisition and transfer.

2.3.2 Skill acquisition as memory retrieval

The Instance theory

The approach in Logan's (1988, 1990) Instance theory to accounting for the phenomena of skill acquisition is very different to that of theories like ACT-R and SOAR. The latter theories generally describe improvement with practice as resulting from the refinement and tuning of procedural knowledge. In contrast, the Instance theory sees improvement in performance, particularly faster performance, as resulting from an increased range of episodic representations of past experience to call upon.

Three assumptions underlie the Instance theory. The first assumption—obligatory encoding—is that the act of paying attention to something results in a memory representation. Thus, each processing episode results in a separate representation in memory. This representation, or instance, consists of the stimulus conditions, the goal state, the interpretation given to the stimulus conditions, the response(s) executed, and the result of the response(s). The second assumption is that of obligatory retrieval. According to this assumption attention to an item or event causes whatever was associated with it in the past to be retrieved from memory. In other words, encountering a set of stimulus conditions that have been encountered again will result in the retrieval of all instances that were stored from those previous encounters. The third assumption, instance representation, says that each episode results in a separate instance, even if the episode is identical to previous ones. Together these assumptions effectively mean that attention to a stimulus will result in the retrieval of all previous associations to that stimulus. There could be many instances being retrieved if experience is extensive, and they all will be the same.

Logan (1988) claims that performance can be described as skilled (automatic) when it relies on the retrieval of instances only. This only occurs after a person has had some experience with a task, and therefore some instances to retrieve. With sufficient practice, exposure to the task will result in the retrieval of a past solution. When performance follows a remembered solution, behaviour is said to be automatic. Clearly, though, when we first encounter a task we do not have any instances to retrieve. So the Instance theory describes a process whereby initial performance with a task can commence, and, with practice, improve.

Development of skill According to Logan (1988), initial performance on a task is under the control of an algorithm. Logan makes no claims as to where these algorithms come from. As far as the other predictions of the theory go, the origin of the algorithm is not crucial. Still, the theory does overlook a substantial number of phenomena associated with developing a skill. Logan (1990) acknowledges the theory's shortcomings on this point and notes that a full understanding of the nature of instances requires some understanding of the algorithms that enable initial performance and result in instance representation.

To say that we perform a task initially by applying an algorithm means that somehow we acquire or develop a method for generating a solution to the task. For example, when children first learn to add single digit numbers, they typically generate an answer by counting (e.g. they might count out each addend on their fingers and then count the total) (Siegler and Jenkins, 1989).

Although initial performance proceeds by the execution of an algorithm, each time the task is executed, something is remembered about that processing episode (obligatory encoding assumption). What is stored in memory is an instance. Each episode results in a separate instance, even if the episode is identical to previous ones (instance representation assumption). In the single digit addition example, each time 2 + 3 is evaluated with the counting algorithm, a representation of the problem ('what's 2 + 3?') and the solution ('5') would be stored in memory. As the task is practiced, more instances are represented in memory.

For any particular processing episode, the algorithm can be executed, and the episode can access instances (obligatory retrieval assumption). In the single-digit addition, for example, 2 + 3 can be evaluated using the tried-and-true counting method, but at the same time, the present episode will retrieve past attempts at solving this expression (i.e. past solutions). As a result, performance in the present episode can take advantage of this retrieval of a past solution. Certainly, children appear to eventually remember the sums of all pairs of single digits, and perform addition by memory retrieval (Siegler, 1987). Thus, with practice, there are two options for controlling performance: algorithmic processing or instance retrieval (i.e. retrieval of a past solution).

According to the Instance theory, only one mechanism can control processing in any one episode. Initially performance is controlled exclusively by execution of an algorithm. As the number of instances increase with practice, a race develops between algorithmic processing and instance retrieval. In any particular processing episode, the two processes will proceed in parallel. Initially the algorithm wins all the time, because there are very few instances to draw from. Those that exist will be retrieved too late to control performance because the algorithm will have been executed already. With practice, more instances are stored. Logan (1988) says that instances have a distribution of retrieval times associated with them. Logan shows that as the number of instances increases, the time it takes to retrieve an instance decreases. Coupled with the claim that algorithm execution time does not change with practice, this means that, as the number of instances increases, the chance of retrieving an instance before the algorithm is executed also increases. Thus, practice leads to a reduction in performance time.

Logan demonstrates how the reduction in performance time conforms to the power law of learning. As practice leads to an increase in the number of instances, the size of the distribution increases. Logan suggests that the distribution of retrieval times is a Weibull distribution. Weibull distributions, and others that are roughly bell-shaped (e.g. the Normal distribution) have a characteristic relationship between an increase in their size and the values of their extremes. Logan has demonstrated how an increase in size is associated with an increase in value of the extremes. This relationship is a power function. That is, the value of the extremes increases as a power function of the size of the distribution. In terms of practice and retrieval times, the shortest retrieval time (i.e. the instance that will control performance) is a power function of the number of instances (i.e. practice). Logan (1990, pp. 5–6) provides a more intuitive description of how performance time on a task improves with practice according to the power law of learning:

> the power function speed-up follows from two opposing tendencies that govern the race: On the one hand, there are more opportunities to observe an extreme value as sample size increases, so the expected value of the minimum will decrease as the number of traces in memory increases. On the other hand, the more extreme the value, the lower the likelihood of sampling a value that is even more extreme, so adding traces to the race will produce diminishing returns. The first factor produces the speed-up; the second produces the negative acceleration that is characteristic of power functions.

Logan has also demonstrated that this relationship between distribution size and instance retrieval time predicts that the standard deviation of performance time will decrease as a power function of practice. Moreover, this power function will have the same learning rate as the power function describing improvement in performance time. Both of these predictions have been supported by experimental data (e.g. Logan 1988).

One of the tasks Logan has used in his demonstrations of features of the Instance theory has been the alphabet arithmetic task. This involves subjects evaluating expressions, such as

$$A + 3 = D \qquad C + 2 = G$$

Subjects are simply required to say whether the statements are true or false (i.e. the first example is true and the second is false).

Logan (1988) claims that subjects initially perform this task by counting through the alphabet. Eventually though, after much practice, subjects invariably perform by remembering solutions to expressions experienced before. Logan (1988) showed that RT on this task and standard deviations of RT were reduced as a function of practice, and that these functions had the same learning rate.

In many respects, Logan's Instance theory is a more elegant theory than most of the theories that describe skill acquisition as a process of strategy refinement. What this means is that Logan proposes a fairly simple learning mechanism (the accumulation of experience) and this accounts for several learning phenomena almost as a by-product. On the other hand, theories, such as Anderson's ACT-R theory, involve several mechanisms designed to account for several phenomena. This gives the theories a bit of an *ad hoc* flavour. Still, elegance is not the only criterion by which theories are evaluated. Accounting for all the data is the most important. And from this respect, the Instance theory does suffer from a few problems, particularly, with respect to the issue of transfer.

Transfer One curious feature of the Instance theory concerns its predictions regarding transfer. Basically, the theory predicts that there is no transfer between different tasks. The theory holds that skilled performance is based on the retrieval of instances. Thus, when we perform a new task, there should be no instances to call upon, and so performance should be at beginner level. Furthermore, since algorithms do not improve with practice in the Instance theory, performing the same task with different items will not be associated with any benefit. This can be illustrated with respect to the alphabet arithmetic task.

Logan and Klapp (1991) had participants practice the alphabet arithmetic task with letters from only half of the alphabet (i.e. half of the subjects practiced with A–M, the other half with N–Z). After 12 sessions of practice, with each problem repeated 6 times per session, participants were then presented with problems involving letters from the other (unpractised) half of the alphabet.

According to the Instance theory, when subjects acquire skill at alphabet arithmetic, they are relying more and more on the retrieval of solutions to particular problems. For example, with the problem A + 3 = D, rather than counting through the alphabet each time to determine 'true' or 'false', participants remember 'true'. In other words, when they see the problem A + 3 = D, they retrieve all the instances associated with that item. This is also the case for all the problems experienced. Each problem has its own distribution of instances to call on. Thus, participants are *not* getting better at counting through the alphabet. As a result, when the items are changed, to ones with letters from the other half of the alphabet, participants no longer have any instances to call upon. So their performance should return to beginner levels. This is indeed what Logan and Klapp found. When participants were exposed to the transfer items, the gains achieved during training were lost. This result supports Logan's (1988) claim that skills are item-specific, that is, the skills acquired are specific to the items experienced during training and cannot be used to perform the task with other items.

Although Logan's experiments have supported his claims about transfer being item-specific, on the whole his transfer prescription is too limited. That is, transfer has been observed by others to be far more general than Logan claims. Many researchers (see Chapter 3) have reported evidence that is consistent with general algorithms being developed that can be applied to many different items. Even the alphabet arithmetic experiments (Logan and Klapp, 1991) reveal some evidence that, although disrupted considerably by the change in items, participants did develop some item-general skill (i.e. they did not return to beginner levels—so they were getting better at counting through the alphabet). Other results that Logan has reported as evidence for item-specific skills have also indicated the presence of item-general skills (e.g. lexical decision (Logan, 1988, 1990); dot counting (Lassaline and Logan, 1993); detection of words in particular categories (Logan and Etherton, 1994)).

Although item-specific information seems to be represented, and restricts transfer, item-general information is also represented, sometimes. It is possible that Logan has found item-specific transfer because of the highly constrained tasks he had his research participants practice. When task conditions are different (e.g. Speelman and Kirsner, 1997), item-general transfer occurs. Speelman and Kirsner suggested that it is task conditions that determine the nature of skill and transfer, not the inherent nature of skill. That is, skills do not have a fixed nature, but rather develop as an adaptation to the environment. This issue is considered further in Chapter 3.

Mounting evidence that the Instance theory is too restrictive in its transfer prescriptions has prompted some suggestions for modifying the theory. Logan (Logan *et al.*, 1996, 1999), for example, has suggested that instances may be retrieved in a piecemeal fashion, that they may possess several parts that can be retrieved separately. Furthermore, there may be differences in the rate at which particular parts of an instance can be retrieved. In this way, the Instance theory could account for situations in which people appear able to perform tasks that are not identical to practiced tasks, and yet their performance exhibits some benefit of the prior practice. Importantly, the extent of transfer in such situations appears to be a function of the similarity between the two tasks. Thus, any modification to the Instance theory would need to include a mechanism whereby task similarity can determine the extent of transfer.

One modification to the Instance theory that provides a means of accounting for similarity-based transfer has been proposed by Palmeri (1997). His exemplar-based random walk (EBRW) model retains the race between an algorithm and memory retrieval. The major change to the theory in the EBRW model, however, involves a retrieval mechanism that is sensitive to similarity differences between instances. That is, experiencing an event can result in the retrieval of instances that are only similar to the event, rather than identical (as in the original theory). In addition, the speed with which an instance is retrieved is inversely proportional to the degree of similarity that exists between the event and the instance. Thus, the greater the similarity, the faster the retrieval time, with identical instances being retrieved the fastest (in such a case the EBRW model's predictions match those of the instance theory). By incorporating similarity-based retrieval of instances, the EBRW model allows for transfer between non-identical experiences (i.e. learning is not always as item-specific as implied by the Instance theory). This loosening of the boundaries on transfer, though, comes at a cost. Unless experiences are identical, transfer will be less than complete and be a function of similarity.

2.4 **Conclusions**

The history of research in skill acquisition is long and full of a vast amount of tinkering with variables to gauge their effects on learning and performance. In spite of the huge amount of information that has been collected on these issues, two generalizations are possible: (1) practice leads to performance improvements that are dramatic early in practice and diminish with further practice; and (2) transfer between tasks is a function of the degree of similarity between the two tasks. Both of these generalizations are, of course, subject to qualifications, and yet, they both form the basic assumptions of most theories of skill acquisition. As indicated in this chapter, however, the theories account for these basic 'facts' of skill acquisition to varying degrees. Although all of the theories presented provide an account of the power law, with respect to the issue of transfer there is greater unevenness. Crossman's theory relied on Thorndike's identical elements model and so could be considered to be limited in the same manner as that model (i.e. transfer between tasks is a function of similarities in observable stimulus–response elements only). The ACT-R and SOAR theories, as well as MacKay's theory, all account well for the transfer effects presented so far. The Connectionist theories of skill acquisition have not explicitly tackled transfer as yet. Finally, the Instance theory enjoys equivocal success on the transfer issue because it predicts transfer to be highly restricted and this is only observed in certain circumstances. More recent modifications of Logan's theory by others, however, suggest that this shortcoming can be overcome. As is indicated in the following chapter, however, the transfer issue is more complicated than has been presented up to now. For instance, whether transfer is general or specific to the items experienced during training appears to be related to several features of the environment in which performance takes place, not the inherent nature of skills. In addition, transfer has a dramatic effect on the shape of learning curves. We argue then, that it is on this issue that all current theories flounder and hence ultimately fail the Skill Acquisition Phenomena Criterion for suitability as a theory of skill acquisition.

At this point, it is possible to begin to evaluate the accounts provided by the various theories of the power law against the Explanation Criterion mentioned in Chapter 1. Although all of the theories provide an account of the law, some theories rely on building this feature into the theory as an explicit constraint on the way that practice affects performance. So, the ACT-R account of the power law rests in large part on the strengthening mechanism (i.e. compilation may be responsible for speed-up early in practice, whereas strengthening is responsible thereon). That is, the performance speed of a task is determined by the strength of the productions underlying performance of the task. A production's strength is determined by the history of successful application of the production (i.e. practice). ACT-R defines strength as being a power function of practice. And since performance speed is a direct function of strength, performance speed is thus a power function of strength. There is no explanation, however, of why strength is a power function of practice—it just is. Similarly, MacKay's theory builds in a negatively accelerated feature: the priming rate of nodes is an exponential function of the frequency of node activation. All of the other theories considered, however, account for the power law as a by-product of the way that practice leads to changes in the nature of mental representations (Crossman's theory, SOAR, Connectionist theories) or to changes in the number of mental representations (Instance theory). It can be argued then, that the accounts of the power law provided by ACT-R and MacKay are really just re-descriptions of the law. The other theories, however, explain the law as an emergent property of practice effects.