REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
15-06-20003

2. REPORT DATE
Final Technical Report

3. DATES COVERED *(From - To)*
July 1999 – December 2001

4. TITLE AND SUBTITLE
Feedback and Transfer in the Acquisition
of Cognitive Skills

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-99-1-0929

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHOR(S)
Ohlsson, Stellan

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Illinois at Chicago, Department of Psychology
1007 West Harrison Street
Chicago, IL 60607

8. PERFORMING ORGANIZATION REPORT NUMBER
June 15, 2003

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Ballston Centre Tower One
800 North Quincy Street
Arlington, VA 22217-5660

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
N/A

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The acquisition of cognitive skills requires feedback and support for transfer of training. It was found that negative feedback is optimal when deliverd with narrow scope, while positive feedback is optimal when delivered with wide scope. It was also found that conceptual instruction did not support transfer of training more than practice,when transfer was measured as accuracy. However, the underlying transfer process was nevertheless different, as revealed by a componential analysis of the solution times. The second finding deserved further study.

15. SUBJECT TERMS
Abstraction, cognitive skill, feedback, training, transfer

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Stellan Ohlsson
UU	SAR	SAR	SAR	117	19b. TELEPONE NUMBER *(Include area code)* (312) 996-6643

Feedback and Transfer in the Acquisition of Cognitive Skills

Stellan Ohlsson,
Department of Psychology
University of Illinois at Chicago

Technical Report

June 15, 2003

With Contributions From:

Andrew Corrigan-Halpern
Timothy Nokes

This technical report serves as the Final Technical Report for ONR Grant #N00014-99-1-0929 ("Optimizing the acquisition of hierarchical task representations"). No endorsement should be inferred.

Summary

The acquisition of complex cognitive skills raises two central research problems: The function of feedback during training and the cognitive mechanisms by which knowledge gained during training is transferred to a new task context.

Although feedback has been a topic of research for most of the past century, many findings are based on studies of simple motor learning and provide little guidance for how to provide feedback for the acquisition of more complex skills. Several variables have been shown to impact the effectiveness of feedback, but there is no consensus as to how these variables achieve their effect. Little work to date has considered how feedback operates for skills that are organized hierarchically. We conducted two studies where feedback is given for complex letter extrapolation problems. We introduce a variable that we call scope. Scope refers to whether feedback is given for an individual action (local scope) or for a sequence of actions that serve a common goal (global scope). In our first study we show that scope interacts with the total amount of feedback given. In our second study we present an interaction between feedback scope and feedback type (positive feedback verses negative feedback). We claim that the interactions presented cannot be explained by the currently held theories of feedback and we suggest directions for further research that aim to clarify these effects.

Contemporary theories of learning postulate one or at most a small number of different transfer mechanisms, i.e., processes by which what is learned in one context is projected onto, or applied to, another task context. However, people are capable of mastering a given task via qualitatively different learning paths. We hypothesize that the knowledge acquired via such alternative paths differs with respect to level of abstraction and the balance between declarative and procedural knowledge, both factors that affect transfer. In three experiments, we investigate what is learned about patterned letter sequences via either direct instruction in the relevant patterns, practice in solving letter sequence extrapolation problems, or an incidental learning procedure. Mastery of the target task can be achieved either via instruction or practice, but the structure of the solution times reveals significant differences in cognitive processing. The findings are not in accord with the standard trade-off view of the distinction between declarative and procedural knowledge. Learning theories that claim generality should be tested against cross-scenario phenomena, as opposed to parametric variations of a single learning scenario.

Table of Contents

Section Page

Part I: Feedback in the Acquisition of Cognitive Skill: Reopening the Issue 5
Background 6
General Method 12
Experiment 1 16
Experiment 2 23
General Discussion 25
References 30
Table Captions 34
Figure Captions 36

Part II: Comparing Multiple Paths to Mastery: What is Learned and Transferred? 47
Background 48
Task Environment 54
Overview of Studies 56
Experiment 1: Baseline 57
Experiment 2: Practice versus Direct Instruction 59
Experiment 3: Practice versus Incidental Training 68
Measures of Transfer 74
Towards a Processing Account 76
General Discussion 82
References 87
Appendix A 90
Table Captions 92
Figure Captions 106

Part I:

Feedback in the Acquisition of Cognitive Skill: Reopening the Issue

Background

Like other types of skills, cognitive skills are learned by practicing. For practice to be effective, the learner must receive feedback about his or her performance. Psychological theories of skill acquisition are incomplete unless they explain how learners benefit from feedback. It is of great practical importance to understand how feedback should be crafted to optimize learning. However, the bulk of research on feedback has concerned motor skills, so there are few empirically grounded generalizations about feedback in the acquisition of cognitive skills. In particular, there is little evidence for the widespread belief that positive feedback is more effective than negative feedback. Furthermore, research on feedback has overlooked the fact that cognitive skills are hierarchically organized. We show in two experiments that organizational level interacts with rate and type of feedback

Definition and Focus

Broadly conceived, feedback includes any information that reaches the learner from the environment about the appropriateness, correctness or effectiveness of his or her actions. In some task environments, such information consists of the observable causal consequences of the relevant actions, e.g., the behavior of a device after a particular button has been pressed. This form of feedback is often referred to as intrinsic feedback (Salmoni, Schmidt, & Walter, 1984). In learning and training environments, feedback typically includes instructional discourse delivered by a coach, teacher or tutor (human or artificial). This type of feedback is often referred to as extrinsic feedback (Schmidt, Young, Swinnen, & Shapiro, 1989).

Feedback takes different forms in cognitive (Anderson, 1981; Ohlsson, 1996b) and motor skills (Patrick, 1992; Proctor, 1995; Salmoni et al., 1984). Although the distinction cannot be made perfectly sharp, we regard a skill as primarily motor when task performance depends on the exact physical movements by which the relevant actions are implemented. For example, when striking a ball in tennis, it matters exactly how the racket is swung, with what angle, speed, direction, etc. In contrast, a skill is primarily cognitive when the physical properties of the relevant actions do not matter for task performance. For example, it matters not for the outcome of a chess game by which physical movement a chess piece is moved; the move is the same, whether the piece is moved by hand, foot or mouth. Clearly, feedback about the trajectory of one's hand in moving a chess piece would not be helpful in learning to play chess. Because we focus on cognitive rather than motor skills, we consider only high-level conceptual feedback such as information about the correctness of a problem solution, as opposed to, for example, information about the spatial parameters (speed, location) of a movement in relation to a target movement (Adams, 1971; Salmoni et al., 1984; Welford, 1968).

Background

E. L. Thorndike was the first learning theorist to systematically consider the effects of what he called after-effects on learning (Thorndike, 1913a) On the basis of a large number of empirical studies, he claimed that the after-effects have a stronger impact on the learner than mere repetition. He formulated the influential Law of Effect, which in its initial form claimed that what Thorndike called satisfiers (rewarding after-effects) strengthen the connections between situations and actions, while what he called annoyers (punishing after-effects) weaken such connections (Thorndike, 1913a, 1913b). As the terms "satisfier" and "annoyer" indicate, Thorndike did not distinguish between motivational and informational aspects of after-effects. Other behaviorist learning theorists in the first half of the 20th century discussed the effect of the environment's response to an organism's action in terms of reinforcement (Hilgard & Bower, 1966; Skinner, 1938), a concept that also failed to discriminate between information and motivation. In the terminology of reinforcement, "positive" meant praise or reward (e.g., a food pellet for an animal subject), while "negative" meant punishment (e.g., an electric shock).

The term "feedback" originated in engineering work on servo-mechanisms during World War II (Wiener, 1948), and migrated from control theory into psychology in the late 1940s (Arbib, 1964; Craik, 1947, 1948; Moray, 1987; Murray, 1998). Some psychologists saw the concept of a feedback circle as having the theoretical potential to replace the reflex arc as the building block of nervous systems (Miller, Galanter, & Pribram, 1960). The shift from reinforcement to feedback lead researchers to emphasize the informational rather than the motivational aspect of the environment's response to an action: What does that response tell the learner about his or her current task or situation, and how does he or she make use of that information to improve task performance? Consistent with this orientation, some authors have used the term "knowledge of results" instead of "feedback" (Bilodeau, 1966). From our point of view, there is no conceptual difference between feedback and knowledge of results; we prefer the shorter term.

In engineering, the terms positive and negative feedback have technical definitions that for obvious reasons are devoid of any motivational implications. Briefly put, positive feedback is present when the output of a device is connected to its input in such a way that an increase in output produces further increases in output; negative feedback is present when the connection is such that an increase in output leads to a future decrease in output. Although the informational orientation survived the migration into psychology, these precise definitions did not (Moray, 1987). Within psychology, "positive feedback" have come to mean information that informs a learner that an action or solution is appropriate, correct or effective, while "negative feedback" means information to the effect that an action is inappropriate, incorrect or ineffective. This is how we use these terms in this article.

Throughout the historical changes in conceptualization, there has been consensus that feedback is essential for successful skill acquisition. Given that feedback is helpful, one would expect the learner to benefit more, the higher the feedback rate. If each feedback message provides additional information, then learning should be most effective when the learner receives feedback after 100% of his or her actions. However, in realistic learning scenarios the feedback rate can be considerably lower. Surprisingly, laboratory experiments that systematically vary feedback rate have not consistently confirmed the expected effect. In some studies, increasing the feedback rate improved performance (Kulik & Kulik, 1988; Salmoni et al., 1984; Schmidt et al., 1989; Thorndike, 1927). In others, increasing the rate was found to lower performance (Bourne, 1957; Bourne & Bunderson, 1963; Schroth, 1997). The obvious inference is that the effect of feedback rate is mediated by other variables.

Feedback Type

In the late 1920s, Thorndike revised his Law of Effect to say that satisfiers and annoyers act in different ways. Thorndike (1931; 1927; 1932) claimed that " ... the strengthening of a connection by satisfying consequences seems, in view of our experiments and of certain general considerations, to be more universal, inevitable, and direct than the weakening of a connection by annoying consequences." Although satisfiers directly strengthen the connections between situations and actions, annoyers act primarily by prompting the learner to try other actions, which then might be strengthened instead. The weakening of incorrect connections is indirect. Laboratory findings consistent with the revised law (Thorndike, 1932), the common tendency to confound informational and motivational effects of feedback, and the claim of progressive educators that encouragement leads to better educational outcomes than punishment have combined to produce a widespread belief that psychological research has shown that positive feedback is more effective than negative feedback.

However, the issue of how positive and negative feedback functions in the acquisition of cognitive skills is not, in fact, settled. First, once motivation is distinguished from information, the superiority for praise over punishment does not necessarily imply that positive feedback is more helpful than negative feedback. Second, the early 20th century laboratory studies cannot be regarded as decisive due to methodological weaknesses. For example, Thorndike often compared small groups of subjects without the benefit of statistical significance tests (Thorndike, 1932), and his choice of chance-level base line performances has been criticized (see Hilgard & Bower, 1966). Many of the results from the reinforcement tradition were obtained with animal subjects, and it is not obvious that they can be generalized to the acquisition of complex cognitive skills by humans. Third, most studies of feedback have focused on motor skills (Salmoni et al., 1984). Although it is likely that there are some similarities between the two types of skills in how they are affected by feedback, it is also likely that there are some differences, so results cannot be generalized without further evidence.

Finally, the empirical data base is small. In the period 1930-1980, the issue of positive versus negative feedback did not receive much attention, presumably because it was considered settled. There were many studies of cognitive learning in which feedback was a component of the experimental procedure, but in all but a few feedback was a tool for establishing a desired level of learning rather than the phenomenon being investigated. In an extensive review, D. B. Ausubel (1968) summarized the status of the field as follows:

> On theoretical grounds, knowledge of results or feedback would appear to be an extremely important practice variable. Nevertheless, because of serious gaps and inadequacies in the available research evidence, we possess very little unequivocal information either about its actual effects on learning or about its mechanism of action. (p. 315)

Consistent with this assessment, we have been unable to locate any systematic body of empirical research on the functions of positive and negative feedback in the acquisition of multi-step cognitive skills in the period 1930-1980.

During the 1980s, pioneers in cognitive modeling posed the challenge of simulating the acquisition of cognitive skill (Anderson, 1976; Anderson, 1981; Anderson, 1982; Anzai & Simon, 1979; Newell, 1990). Researchers began to formulate computational mechanisms that were hypothesized to correspond to the cognitive processes underlying skill acquisition (Anderson, 1981). Some of the proposed mechanisms operated independently of feedback. For example, Anderson (1983) hypothesized that units of skill knowledge called production rules become integrated on the basis of an internal record of their temporal succession during action. Ohlsson (1987b) and others (Klahr, Langley, & Neches, 1987) proposed other optimization mechanisms that also operated by analyzing the 'mental code' for a skill rather than environmental feedback. Other mechanisms did depict learning as operating with feedback. For example, Langley (1987) hypothesized that production rules that are sometimes followed by positive and sometimes by negative feedback are replaced by more specific rules that incorporate the conditions that differentiate between the two types of situations. Positive feedback was generally thought to strengthen a production rule, i.e., increase its probability of retrieval as stated in the first half of Thorndike's Law of Effect, but without any change in content (Anderson, 1983). This hypothesis enjoys enduring popularity.

This flowering of theory did not lead to a renewed interest in the different functions of positive and negative feedback. The various learning mechanisms were not typically specified in those terms, and the fact that several hypothesized mechanisms did not rely on any type of feedback was not discussed. The number of empirical studies of positive and negative feedback did not increase.

In our own work on learning from negative feedback (Ohlsson, 1987a, 1993; Ohlsson, Ernst, & Rees, 1992; Ohlsson & Jewett, 1997; Ohlsson & Rees, 1991) we began with the observation that learners often know what a correct problem solution should look like, how a device should behave when operating correctly, and so on. Such expectations can be conceptualized as constraints on correct task solutions, and negative feedback about an error as a signal that this or that constraint has been violated. We developed a computational mechanism by which the information embedded in a constraint violation can be translated into a revision of the violating production rule. The rule is thereby specialized in such a way that it does not become active in situations in which it causes errors, which gives other, more appropriate rules the opportunity to apply. This learning mechanism thus incorporated Thorndike's insight about the indirect operation of negative feedback. It was embedded in a computer simulation called HS which was capable of acquiring skills in elementary arithmetic and chemistry (Ohlsson, 1996a, 1996b; Ohlsson et al., 1992; Ohlsson & Rees, 1991). HS remains the most developed model of learning from negative feedback in the cognitive modeling literature. It has been shown to produce negatively accelerated learning curves (Ohlsson, 1996b), and it has informed the design of empirically successful intelligent tutoring systems (Mitrovic & Ohlsson, 1999). Nevertheless, it, like other computational models of cognitive skill acquisition, is seriously underspecified by data, and many of its predictions remain untested (Ohlsson, 1996b). For present purposes, an additional limitation is that the model did not learn from positive feedback.

Positive and negative information might need to be processed in qualitatively different ways. Positive feedback provides definitive information. If the learner receives information to the effect that action A in situation S is appropriate, then he or she knows exactly what to do in future encounters with that type of situation. The information does not require further interpretation or processing. However, when the learner receives information to the effect that a particular action was an error, the information is incomplete. There remains to figure out why the action was incorrect and (hence) which action to perform instead. Hence, positive feedback provides more information than negative feedback. This argument from information content is frequently put forward as a statement of logic that requires no empirical validation. For example, Hilgard & Bower (1966) wrote:

> There is a logical difference between responding in the intelligent direction to Right and Wrong. The intelligent to response to Right is to do again what was last done. This makes possible immediate rehearsal; the task is clear. The intelligent response to Wrong is to do something different, but what to do is less clear. It is necessary both to remember what not to do and to form some sort of hypothesis as to what to do. (page 32)

The argument implies that qualitatively different cognitive processes are required to learn from positive and negative feedback. It also implies that positive feedback should be more helpful than negative feedback.

Although the argument from information content has the appearance of a logically necessary argument, the implication that positive feedback is more helpful than negative does not, in fact, hold up under all circumstances. The empirical literature supports the idea that positive and negative feedback have qualitatively different effects (Taylor, 1991), but it is once again inconsistent. Greeno (1974) did indeed find that positive feedback was more effective than negative feedback, but Mesch, Farh, and Podsakoff (1994) found that negative feedback produced better performance.

To clarify this issue, we conducted a series of simulations using an abstract computer model (Ohlsson & Jewett, 1997). Abstract computer models are similar to mathematical models in that they only capture the quantitative aspects of a hypothesized process or mechanism. The advantage of abstract models is that they are simple enough to enable extensive simulations to be run, which in turn enables the researcher to systematically vary model parameters and to obtain quantitative regularities. The model conducted a heuristic search and it incorporated the almost universally adopted principle that a rule (connection) is strengthened when it was followed by positive feedback. The model responded to negative feedback by correcting the indicated error, i.e., it responded as if the HS learning mechanism (see above) was operating. We found that in this type of system, negative feedback had a stronger effect on the rate of learning than positive feedback (Ohlsson & Jewett, 1997). The reason is that the correction of an incorrect action cuts off the entire search tree that sprouts from the problem state created by that action, and thus preempts a potentially large number of possible future errors. The argument from information content is not valid unless augmented with particular task environment, a particular performance mechanism, and a specification of how feedback is processed. The situation becomes even more complicated when we factor in the organization of skill knowledge.

Hierarchy

The performance of complex cognitive skills seldom consists of the execution of single actions; it is usually necessary to carefully coordinate temporally extended sequences of actions. Although each individual action might be under the control of a single, modular knowledge element, such as a production rule or a motor schema, the performance as a whole must be under the control of a larger structure. Cognitive models of skill typically conceptualize this larger organization as a hierarchical goal-subgoal tree that decomposes a task into subtasks or components. An everyday example is the decomposition of planning a vacation into its component tasks. Completing a trip might involve transportation to the destination, finding accommodation, and transportation back home. Each of these subtasks obviously decomposes into yet small tasks (e.g., dial the hotel reservation desk). The result is a hierarchical structure in which the top node specifies the task as a whole, the intermediate nodes represent subgoals (component tasks), and the terminal nodes represent actions that are so simple that they can be executed without further decomposition at the cognitive level (e.g., reach for the telephone).

Laboratory studies strongly support the hierarchy hypothesis. Several studies (Anderson, 1993, chap. 6; Egan & Greeno, 1974; Greeno & Simon, 1974; Restle, 1970; Restle & Brown, 1970; Ruiz, 1986; Simon, 1972) have reported that inter-response times in the execution of complex tasks show a scalloped pattern, in which the first action after the retrieval of a subgoal has a longer-than average response time, but the actions dominated by that subgoal have shorter times; when the subgoal has been reached and the next subgoal is to be retrieved, the next inter-response time is once again longer than average. Studies on knowledge representation also suggest a hierarchical organization. For example, semantic memory is organized hierarchically (Collins & Quillian, 1969; Miller, 1998). Finally, Simon (1962) in an oft-quoted essay, argued that complex systems such as skills become more probable if they are constructed in successive stages such that in each stage a few components are assembled into intermediate structures, which then in turn are combined into the complete structure. This type of process generates a hierarchical final product.

Although the two concepts of feedback and hierarchy are often mentioned in the context of skill acquisition, they have rarely been considered together. How is feedback processed in the context of a goal-subgoal hierarchy? More precisely, suppose that action A is dominated by subgoal G_n, which in turn is dominated by subgoal G_{n-1}, and next G_{n-2}, and so on, until we reach top goal G_0. If A is followed by feedback, which goal or goals are affected? Is only the immediately superordinate goal, G_n , affected? Only the top goal, G_0? All goals G_i in between? What are the differences between positive and negative feedback in this respect? Which type of feedback should be expected to be most effective in a hierarchical task representation?

The argument from information content appears even stronger in the case of hierarchical task representations. Positive feedback is as definitive in a hierarchical representation as in a single-level one. To know that an entire subtask has been performed correctly is even more useful than to know that a single action has been done correctly. However, the interpretation problem associated with negative feedback is even more complex in a hierarchical representation. The learner has to figure out not only which prior action or actions were incorrect, but also which, if any, of the subgoals are incorrect and hence should be affected by the feedback. The simple version of the argument from information content predicts that the advantage of positive over negative feedback should be even greater for hierarchical task representations than for single-level representations.

We conducted a series of abstract computer simulations to investigate the interaction between hierarchy and type of feedback (Corrigan-Halpern & Ohlsson, 2001; Ohlsson & Halpern, 1998). We found that our model learned faster if the propagation of negative feedback upwards in the goal hierarchy was conservative, i.e., had most of its effects at the lowest levels in the hierarchy, than when the error information was allowed to propagate all the way to the top goal. The opposite was the case with positive feedback: Learning was slightly more effective when such feedback was allowed to propagated upwards and affect goals at all levels in the goal tree approximately equally. Once again, the argument from information content is at best a partial truth that ignores issues about how feedback is processed. Our simulations suggest that the effectiveness of positive and negative feedback is a function of whether the feedback is local, i.e., refers to the terminal nodes in the goal tree (individual actions), or global, i.e., refers to higher level nodes (subtasks). We refer to this variable as the scope of a feedback message, i.e., how large a component of the target performance does a particular message refer to?

To summarize, information concerning the function of feedback in the acquisition of multi-step cognitive skills is scarce. Many of the early studies of feedback were conducted before modern methodological standards were developed. The majority of studies on feedback concern motor rather than cognitive skills. Many of the process hypotheses proposed in the computational modeling literature do not draw upon feedback, and others have not been tested against empirical data. Furthermore, feedback research has overlooked the fact that cognitive skills are hierarchically organized.

The result is that we have few empirically grounded generalizations about the function of feedback in the acquisition of cognitive skills. In particular, we do not have a clear understanding of the effects of feedback rate, we do not know how feedback interacts with goal hierarchies, and we do not know whether positive and negative feedback are processed differently and, if so, how. The argument from information content is insufficient to settle the latter question, because it ignores how feedback is processed.

We report two experiments that begin to address these issues. In Experiment 1, we aim to verify that the scope of a feedback message, i.e., whether it refers to a single action or a larger component of a task, matters for the impact of that message, and to determine how that impact is modulated by the rate at which feedback is delivered. In Experiment 2, we investigate whether positive and negative feedback are processed differently, and how feedback type interacts with scope. In both experiments, we determine the impact of the relevant variables during the learning period, on the subsequent performance of the target task, and on performance on a near transfer task.

General Method

The experimental study of feedback poses multiple methodological challenges. There is little principled discussion of them in the literature. The purpose of this section is to motivate our approach.

Target Task

The class of complex cognitive skills includes algebra, card games and software use. However, ecologically valid tasks tend to be irregular, draw upon unspecifiable parts of the learner's prior knowledge, and be learned under conditions that hinder precise control and measurement of relevant variables. One the other hand, many laboratory tasks used to study feedback are either perceptual-motor tasks or reproductive memory tasks, and we cannot assume that findings from such tasks generalize to complex cognitive skills. The ideal target task shares key properties of ecological cognitive skills but enables data capture.

Our target task is a version of the sequence extrapolation task studied previously (Kotovsky & Simon, 1973; Simon, 1972; Simon & Kotovsky, 1963; Simon & Sumner, 1968). The learner is given a series of letters that exemplify a pattern and he or she is asked to continue the pattern. To solve such a problem, the subject must first identify the pattern in the given sequence and then use that pattern to generate the continuation of the sequence.

In the original version of this task (Thurstone & Thurstone, 1941), the subject was asked to add a single next element to the sequence. In our version, the subject is asked to extrapolate the entire sequence. In addition, we asked participants to extrapolate the sequence from an arbitrarily chosen single letter. For example, to extrapolate the sequence FABFCD beginning with the letter T, the subject would produce TOPTQR. Although the solution to this example is obvious, sequence extrapolation problems can be quite difficult. The reader might want to extrapolate MKNPPNKMNLOQQOLN beginning with the letter T.

In order to track mastery, we imposed a time limit on each extrapolation attempt. The subject was allowed to study the given part of the sequence for 20 seconds, then made an attempt to generate the solution, and finally received feedback. The feedback was available for 45 seconds, then the next trial began. The study time was designed to be too short for any subject to solve any of the experimental problems in a single trial. Unlike the problem solving scenario studied by Kotovsky and Simon (1973) and Simon and Kotovsky (1963), mastery in our scenario developed over a sequence of trials, where each trial contained both study and extrapolation.

Finally, to increase the difficulty of the task, we gave the subjects different sequences to study on every trial. Each study sequence contained different letters but they all exemplified the same pattern. Unlike the subjects in the serial learning studies by Restle (1970) and Restle and Brown (1970), our subjects could not master the target task via memorization. They were forced to go beyond the specific letters and look for the underlying pattern. However, extrapolation began with the same initial letter in every trial. In short, the subject was trying to construct a particular sequence of letters, but he or she had to extract information from multiple given sequences to figure out which sequence was required.

This task has several theoretically interesting properties that make it suitable as a model of complex cognitive skills, in the sense in which medical researchers speak about animal models of human diseases. (a) Prior knowledge. Like most ecological tasks, the process of mastering a letter sequence extrapolation problem draws upon the subject's prior knowledge. However, the prior knowledge, namely the letters of the alphabet, is relatively circumscribed. (b) Conceptual content. Detecting the pattern in the given sequence is potentially facilitated by knowledge of the pattern-forming operations that were used to generate the given sequences (see below). The operations constitute a kind of theory of the relevant type of pattern, and solving a problem involves applying that theory. (c) Levels of abstraction. The pattern in a particular sequence can be encoded either concretely, with reference to the specific letters that appear in the sequence, or abstractly, with reference only to the relations between the positions in the given sequence (Nadel, 1991; Nokes & ohlsson, 2000; Ohlsson, 1993). (d) Complexity. To solve a sequence extrapolation problem, the subject must generate a series of coordinated responses, rather than the single responses required in recognition and classification tasks. (e) Generativity. A sequence extrapolation problem cannot be solved by memorizing and recalling the given sequence, but requires complex inferences ("this letter corresponds to that letter over there, and that letter has such and such a relation to this other letter, so the letter in this position should probably have that same relation to this letter, and therefore it must be ..."). (f) Hierarchy. The patterns have meaningful components, which increases the probability that the subjects encode them hierarchically. The components arise as a side effect of the pattern forming operations used. For example, reflection of the subsequence XYZ produces XYZZYX which has the two subsequences XYZ and ZYX. Restle (1970) and Restle and Brown (1970) found strong evidence for hierarchical encoding in their serial learning task. The properties (a) through (f) circumscribe the class of learning scenarios to which one might attempt to generalize our findings.

The given sequences were presented on a computer screen, and the subjects extrapolated them by clicking on empty boxes that symbolized the N places in the sequence and then typing in the letter they thought was the right one for that place (see below for more details). The computer was essential for the definition of both our independent and dependent variables.

Independent Variables

Feedback Type. Learning environments designed for instruction and training often provide feedback with rich explanatory content. Such messages are difficult to equate with respect to meaningfulness and utility for the learner. In the present studies, we provided only basic feedback about correctness. As in Thorndike's early studies, positive feedback consisted of the word "Correct" appearing on the computer screen in connection with a response and negative feedback consisted of the word "Wrong." Both types of messages could refer either to a single response (letter) or to a larger segment of behavior (see below for details).

Feedback Probability. To separate the effects of positive and negative feedback, we decided to vary the total amount of the two types of feedback independently of each other. This approach addresses the following conceptual difficulty: In task environments in which every action can be classified as either correct or incorrect, the learner is able to make inferences from the absence as well as from the presence of feedback. If 100% negative feedback is provided, then the learner can infer that any response that is not followed by feedback is correct (and vice versa). That is, such a learning environment implicitly provides 100% positive feedback as well. This makes it impossible to separate the effects of positive and negative feedback.

In order to vary the amount of positive and negative feedback, we provided feedback probabilistically. That is, we defined experimental conditions in terms of the probability that the subject will receive feedback of a particular type, given that he or she produced a response for which that type of feedback is appropriate. Seventy-five percent negative feedback thus means that there is a 75% probability that the subject will receive feedback after an erroneous response (i.e., writing a particular letter in a particular position in the sequence). This allows us to deliver a single type of feedback at varying densities. This variable can also be varied independently for each type of feedback. An experimental condition might hence be defined as delivering 40% negative feedback and 90% positive feedback. We refer to a specification of this sort as a feedback regimen.

The second component of our approach is to introduce an explicit "no feedback" signal. Our subjects were instructed that the appearance on the computer screen of the word "None" in connection with a response means that there is no feedback, i.e., that the response is either right or wrong; the computer system is not telling them which is the case. This feature was intended to lower the subjects' tendency to interpret the absence of feedback.

Scope. If skills are hierarchically organized, it is likely to matter at what point in the hierarchy feedback is applied. The methodological problem is that the higher-level nodes (subgoals) in the subject's representation of the task are inside the subject's head, loosely speaking, and hence not available as shared referents for instructional discourse. A learning environment or a tutor cannot meaningfully say, "There is something wrong with your conception of subgoal G_{0023}." The only shared components of a problem solution are the terminal nodes in the hierarchy -- the individual actions -- and feedback always arrives after some action or another. Without additional information, the learner does not know to which prior action or actions the feedback message refers.

Our technique for gaining experimental control of hierarchical level utilizes the fact that a high level node in the task representation is likely to correspond to a meaningful component of the task. We use an explicit scope marker, a line that underscores a portion of the extrapolated sequence. This marker informs the learner which letter or letters in his or her extrapolation a feedback message refers to. In the applications reported here, this can be either a single letter (local feedback) or a set of four letters that together constitute a component of the relevant pattern (global feedback). Positive local feedback thus means "this letter is right" and negative local feedback means "this letter is wrong." In contrast, positive global feedback means "all four of these letter are correct" while negative global feedback means "at least one of the these four letters is incorrect."

It is important to distinguish scope from delay, a variable that we do not consider in this article. If feedback is delivered immediately after a response, then it is immediate. It might seem that feedback that awaits the completion of a sequence of responses must necessarily be delivered later, and hence that scope and delay are necessarily confounded. Our approach to this problem is to wait to deliver all types of feedback until the subject declares himself or herself finished with an extrapolation attempt. Then all feedback messages are presented at once and in parallel. With this procedure, there is no relation between scope and delay.

Dependent Measures. There are at least three different potential effects of feedback, each requiring a different evaluation. First, feedback might primarily affect the skill acquisition process itself. That is, the efficiency of the learning mechanisms responsible for skill acquisition might depending on the feedback available in the task environment. If the impact of a particular feedback regimen is beneficial, it ought to increase the learning rate. This type of effect can be evaluated by exhibiting learning curves that track quality or speed of solution attempts across trials.

Second, learners who receive a particular feedback regimen might encode their knowledge of the task differently. If they develop a better or more powerful representation, then feedback ought to affect their performance of the target task after training. Effects on learning and performance are logically and operationally distinct. It is possible that some feedback regimens speed up learning, i.e., that a learner will reach a given performance level in less time or fewer trials, but that two learners who have reached the same performance level, by whatever routes, will perform similarly after training. However, if feedback impacts the task representation rather than (or in addition to) the process by which it is acquired, then the two learners might perform differently as well.

Third, if feedback regimen influences the learner's encoding of the task knowledge, it might affect how well or to what extent he or she can apply that knowledge to transfer tasks. In this case, a transfer task is a sequence extrapolation problem in which the subjects must extrapolate a new instantiation of the pattern, that is they must generate a new sequence of letters that conform to the relations given during the learning stage. The direction of the impact of feedback on transfer cannot be determined a priori. A more abstract task representation might increase the amount of transfer, but a more efficiently compiled representation might be more task specific and hence have the opposite effect.

Learning, performance and transfer are distinct dimensions of the impact of feedback. In the past, different studies have measured the impact of feedback in different ways. Studies that have used multiple measures have found that the most effective feedback condition as measured during learning is not always the most effective condition as measured during transfer (Schmidt et al., 1989; Schroth, 1997). Since there is no reason to favor one measure over another, our approach is to capture data on all three types of potential effects.

Experiment 1

The purpose of Experiment 1 was to determine the effects of probability and scope of feedback on learning in our modified version of the letter sequence extrapolation task. Common sense would lead us to expect that if feedback helps, then increasing the probability of feedback should boost either learning, performance, transfer, or, possibly, all three. Half the subjects in Experiment 1 received 25% feedback, i.e., the probability of receiving feedback after a response was 25%, and the other half received 75% feedback.

We did not vary feedback type. All subjects received both positive and negative feedback, as would be the case in most real learning scenarios. Furthermore, they received both types of feedback at the same probability. That is, those who received 25% positive feedback also received 25% negative feedback; similarly for the 75% condition.

The second purpose of Experiment 1 was to determine whether scope affects behavior in our learning scenario. Scope, as distinguished from delay, has not been systematically studied in prior work on feedback. Half the subjects received only local feedback and half received only global feedback. We made no specific predictions about the effects of scope.

Method

Participants. The subjects were 104 undergraduate students enrolled in the introductory psychology course at the University of Illinois at Chicago. They received course credit in return for their participation. There were approximately 60% females and 40% males. No other demographic data were collected about the participants.

Materials. There were 12 letter sequences with 16 letters each that served as the given strings. All 12 sequences instantiated the same pattern. The pattern is shown in Figure 1. The first half of the pattern consists of an initial group of four letters, followed by a second group that is a reflection of the first. The initial four letters are related to each other via repeated next relations, i.e. one can move from letter to letter by moving either forwards or backwards in the alphabet. The second half of the pattern repeats this two-group structure, but also moves it one step forward in the alphabet. The pattern thus consists of a three-level hierarchical structure with the top level dividing the total sequence into two subsequences of eight letters each. These are in turn divided at the next level into two shorter subsequences. The latter consist of four letters each.

The first 12 trials were learning trials, where subjects viewed a sequence and then reproduced it. They were followed by two assessment trials and two transfer trials. The subjects' task was always to extrapolate the sequence. The first letter in the extrapolation was given as a prompt and it was the letter M on the 12 learning trials and on both assessment trials. The subjects generated the letters for the following 15 positions. In the transfer trials, the initial prompt was the letter T instead. To answer a transfer task, the subject thus had to generate a sequence of letters that he or she had not generated before.

Table 1 contains the 12 given strings, the prompts, and the correct 15-place extrapolations. There was no overlap between the given sequences and the correct extrapolations. That is, the extrapolations could not be constructed by memorizing the given sequences.

Design. Subjects were randomly assigned to four groups, created by crossing feedback probability (low-probability verses high-probability) with scope (local-scope versus global-scope). If a subject was assigned to a low (high) rate group, then he or she had a 25% (75%) probability of receiving feedback after a response. The subjects in the local scope groups received feedback that referred to individual letters, while the subjects in the global scope groups received feedback that referred to groups of four letters. Following the structure of the pattern, the groups consisted of the letters 1-4, 5-8, 9-12, or 13-16.

Procedure. The procedure was the same for all four experimental groups. It consisted of instruction, learning, assessment and transfer. The subjects sat in front of a desktop computer with a 15-inch monitor. All experimental materials were presented via the PsyScope experiment controlling software, which also recorded the subjects' responses and response times.

(a) Instruction. The subjects were told about the three pattern forming operations that were used to construct the sequential pattern underlying the 12 letter sequences. First, the software described the Next operation ("One way to make a letter pattern is to go forward or backward in the alphabet."). They were shown two examples sequences, one going forwards [A B C D E] and one backwards [E D C B A]. They were then asked to extrapolate a simple sequence that followed the rule (in this case one going forwards).

They were given the prompt [P Q] and extrapolated three additional letters. The correct answer was thus [P Q R S T].

Second, the subjects were instructed in the Reflect operation ("Another way to make a pattern is to reverse letters."). They were shown the sequence [A B C C B A] as an example of a reversal, and asked to extrapolate the prompt [J K L] to three places, thus the correct answer was [J K L L K J].

Finally, the subjects were instructed in the Repeat operation ("The third way to make a letter pattern is to repeat letters."), and extrapolated the given sequence [A B C A B C] from the prompt [L M N] to form the sequence [L M N L M N].

These exercises served the two purposes. First they provided instruction on the rules just described. Second, they familiarized the subjects with the study-extrapolate format of the subsequent trials, as well as with the idea of beginning the extrapolation from a given prompt or starting point. Exit interviews revealed that they were effective in both regards.

(b) Learning. The twelve learning trials all followed the same format. The given letter sequence was presented on the screen with the letters in a horizontal row, for a total of 20 seconds. The PsyScope program sampled randomly without replacement from the stock of 12 given sequences (see Table 1), thus producing a different random order of presentation for each subject.

After the 20 seconds of study time, the given sequence was replaced by a single letter followed by a sequence of 15 boxes, also in a horizontal row. This letter will be referred to as the prompt. The prompt showed subjects where to begin the extrapolation. The prompt was the same on every learning trial; hence, the correct extrapolation was the same sequence of letters on each learning trial. The subjects extrapolated by filling in the 15 empty boxes to the right of the M. A box was filled by clicking on the box with the mouse and then pressing the relevant key on the keyboard. The corresponding letter appeared in the box. The subjects could fill in the boxes in any order. They could revise the content of a box by clicking on the filled box and typing a different letter; that letter then replaced the previous letter; the previous letter was no longer visible. The subjects filled in as many boxes as they could, and then clicked on the word "Done" on the screen. They could spend as much time as they wished on the extrapolation task.

Once the subject clicked the "Done" button, the feedback appeared on the screen. The feedback consisted of the words "Correct" and "Wrong", although "Correct" was abbreviated to "Corr" due to space limitations on the screen. Subjects received either local or global feedback according to experimental condition. Colored lines were used to indicate scope. For example, positive local feedback was signaled by underlining the relevant letter with a green line under which the abbreviation "Corr" appeared in green font. Global negative feedback was signaled by underlining the relevant group of four letters with a red line, under which the word "Wrong" appeared in red type.

In the local feedback condition, letters for which feedback was not supplied were underlined with a white line, under which the word "None" appeared in white font. In the global feedback condition, groups of four letters for which feedback was not supplied were underlined with a white line, under which the word "None" appeared in white letters.

The feedback remained on the screen for 45 seconds. When the feedback disappeared, the next learning trial began with the presentation of a different given sequence of 16 letters. There were 12 learning trials.

(c) Assessment. After the twelfth learning trial, there were two assessment trials in which the subjects were asked to produce the target sequence without any further opportunity to study any example of the relevant pattern. They were not shown any study sequence. The program went directly to the extrapolation screen. The subjects tried to fill in the 15 empty boxes, starting with the prompt "M", as they had done in the learning trials. No feedback was given. The two assessment trials were identical and subjects could spend as much time as they desired on each trial.

(d) Transfer. After the two assessment trials, there were two transfer trials. The subjects were not shown any study sequences. They once again produced a 15-item letter sequence that exemplified the same pattern as in the learning and assessment trials. However, instead of being asked to begin the extrapolation with the letter "M", they were now prompted with the letter "T". They were thus asked to generate a letter sequence that they had never generated before. There was no time limit on this task. No feedback was given during the transfer trials. The two transfer trials were identical.

Once the subject had completed all 16 trials (12 learning, 2 assessment, and 2 transfer), they were debriefed as to the purpose of the experiment and thanked for their participation.

Results

Verification of Hierarchical Task Representation. Response times have been used to reliably demonstrate that knowledge is represented hierarchically (Anderson, 1993, chap 1.6; Collins & Quillian, 1969; Egan & Greeno, 1974; Greeno & Simon, 1974; Restle, 1970; Restle & Brown, 1970; Ruiz, 1986; Simon, 1972). We computed response times for each position of the pattern. We analyzed the last four learning trials, so that we could capture subjects' final representation of its structure. We expected that by this point subjects would represent the pattern hierarchically. Four subjects were removed from the analysis because they omitted responses on one or more of the trials. Figure 2 shows the result.

Subjects spent the longest times at the beginning of the pattern. The first position was the prompt "M", so there are no times for this letter. The next three positions "K", "N" and "P" correspond to the first chunk. The high latency for the first "K" probably reflected initial time planning to reproduce the pattern. The next four positions (5-8) correspond to the second chunk, "PNKM". Subjects could reproduce these letters by reflecting the first chunk and response times were shorter. Response times for these four responses approximate a horizontal line. The next chunk "NLOQ" involves the translation of the first chunk by one letter in the alphabet. This required more effort as indicated by the increase in latency from position 8 to position 9The last chunk, "QOLN", could also be completed by reflectionLatencies for the last chunk match those of the second chunk and the horizontal trend is again present.

Learning. Performance as a function of learning trials is shown in Figure 3A mixed ANOVA was performed to verify that the subjects did indeed learn. All 12 learning trials were entered as the within-subject factor. Number of letters correct per trial was the dependent measure. Feedback scope (local or global) and feedback probability (25% or 75%) were entered as between-subjects measures. Number of letters correct per trial was entered as the dependent measure. The learning effect was significant, F (1100, 11) = 56.89, p < .001. There was no main effect of feedback scope, F (100 , 1) = 0.00, p >

.05. There was no main effect of feedback probability, F (100 , 1) = 0.07, p > .05. Other than the learning effect, no effects involving the repeated measure were significant.

There was an interaction between feedback scope and probability, F (100 , 1) = 4.78, p < .05. This is due to the fact the 25% global and the 75% local groups outperformed the 25% local and the 75% global groups, F (100, 1) = 4.76, p < .05.

Assessment. A mixed within-subjects ANOVA was performed to consider how feedback influenced performance on the assessment trials. Both assessment trials were entered as within-subjects measure. Feedback scope (local or global) and feedback probability (25% or 75%) were entered as between-subjects measures. The dependent measure was the number of letters correct per trial. Figure 4 shows the effect of feedback during the assessment trials. There was a significant effect of trial, F (100 , 1) = 6.67, p < .05, as indicated by the fact that subjects performed worse on the second assessment trial than on the first. There was no main effect of probability, F (100 , 1) = 0.47, p > .05. There was no main effect of scope, F (100 , 1) = 0.04, p > .05.

As was seen for the learning trials, the interaction between scope and probability was significant, F (100 , 1) = 6.70, p < .05. This interaction was again due that the 25% global and the 75% local groups outperformed the 25% local and the 75% global groups, F (100, 1) = 6.69, p < .025.

Transfer. We evaluated the transfer of skill in two ways. First, we performed an ANOVA to test whether feedback influenced performance. Both transfer trials were entered as within-subjects measure. Feedback scope (local or global) and feedback probability (25% or 75%) were entered as between-subjects measure. The dependent measure was the number of letters correct per trial. Figure 5 shows the results for the transfer trials. There was no main effect of feedback probability for the transfer problems of the first problem, F (100 , 1) = 0.22, p > .05. There was no main effect of feedback scope, F (100 , 1) = 0.08, p > .05. The interaction between probability and scope was not significant in this case, F (100 , 1) = 1.43, p > .05. There was no effect of trial, F (100 , 1) = 0.44, p > .05 and none of the other effects involving the repeated measure were significant.

A second way to measure transfer is by considering whether subjects improved from learning to transfer. A comparison of figure 4 to figure 5 suggests that subjects performed as well on transfer as they did on assessment. To test this hypothesis, we entered stage (average performance on assessment verses average performance on transfer) as repeated measures in an ANOVA. Feedback probability and scope were entered as between groups measures. There was no main effect of problem stage, F (100 , 1) = 2.21. There was a significant interaction between scope, probability and stage, F (100 , 1) = 8.46, p < .005. This interaction is seen more easily if we examine difference scores.

A difference score was computed by subtracting the average number of letters correct in the assessment trials from the average number of letters correct in the transfer trials. Figure 6 shows the results. While both the 25% global and the 75% local groups show a performance drop (together they complete 1.14 fewer letters correct during transfer), the 25% local and the 75% global groups show a performance gain (they complete 0.37 more letters during transfer). The difference between these means is significant, F (100 , 1) = 8.65, p < .005.

Discussion

The same pattern of results is shown during all three stages of the experiment. During learning and assessment, there was a significant interaction between feedback scope and probability. This trend was present in the transfer trials, but did not reach statistical significance. This data shows that increasing the probability of feedback did not, on the average, increase the subjects' learning probability, their target performance, or their transfer performance. This result is counterintuitive but in accordance with the fact that conflicting effects of increased feedback rate have been reported in the literature performance (Kulik & Kulik, 1988; Salmoni et al., 1984; Schmidt et al., 1989; Thorndike, 1927).

We also found no main effect of scope. Local feedback depended on the probability at which feedback was given. When feedback was local, increasing feedback probability did indeed increase performance. This is in accord with the common sense expectation that if feedback is helpful, then more feedback should be more so.

However, increasing global feedback had the opposite effect on learning, suggesting that an interpretation in terms of amount of information is not sufficient. When feedback was global, increasing the probability of feedback led to lower performance. One possible explanation for this effect is that subjects had difficulty interpreting the global feedback. For the 75% feedback probability condition, there would have been an average of three feedback messages per trial. It may be that subjects in these groups could not process all three messages, and the attempt to do so might have interfered with their learning. Complex cognitive processes, such as the execution of problem solving strategies, can take up enough cognitive capacity to interfere with the acquisition of skill knowledge (Sweller, 1994; Sweller & Chandler, 1994). Analogously, the attempt to gain something from multiple feedback messages, each of which requires interpretation, might have interfered with rather than helped the acquisition of knowledge about the underlying pattern.

This hypothesis is supported by the fact that subjects in the global condition received more negative than positive feedback. In local conditions, 57% of the feedback message came in the form of negative feedback. In the global conditions, 72% of the feedback was negative feedback. A subject received negative feedback for a given four-letter group if at least one of the letters in that group was incorrect. In contrast, he or she received positive feedback for a given group only if all four letters were correct. In the beginning of learning, a subject is therefore more likely to receive negative than positive feedback. As we pointed out in the introduction, negative feedback poses difficult problems of interpretation because the learner must figure out what is wrong with his or her response and what the right response might be. These problems are harder for global than for local feedback. It is possible that the observed interaction between scope and probability is driven by the cognitive load imposed by global negative feedback in conjunction with an interaction between scope and type of feedback. We explore this possibility in Experiment 2.

We found the interaction between probability and scope in the assessment trials as well. There were no study opportunities during the assessment trials and hence no opportunity to acquire additional information about the relevant pattern. The persistence of the differences between the four feedback groups during assessment strengthens the conclusion that the four groups had different knowledge of the target sequence at the end of the learning phase. However, assessment performance does not tell us whether feedback regimen influenced the learning mechanisms (e.g., their speed of operation), the mental representation constructed during learning, or both.

The mental representation might have varied along several dimensions, including level of abstraction. A subject could succeed in the learning and assessment tasks by extracting the underlying pattern from the given sequences, and then attempting to generate the correct extrapolation by filling in the pattern on each extrapolation attempt. In a pure case of this strategy, the subject's representation of the target task would consist of an abstract representation of the pattern plus a procedure for generating an instance of it. We will refer to this as the relation-oriented strategy, because what is learned is primarily the pattern, the relations between positions in the sequence, rather than the letters themselves. On the other hand, a subject could succeed by generating letters via trial and error during extrapolation and use the feedback to decide which were wrong and which were correct. On each trial, he or she would vary those that were wrong and retain the others. In a pure case of this strategy, the subject's representation would be highly specific, consisting of the 15 positions, information of the correct letter for each when known, and lists of incorrect letters for the others. We will refer to this as the letter-oriented strategy, because what is learned is primarily which letter goes in which position. It is likely that the subjects used some mixture of these two strategies; the empirical question is which mixture.

In the transfer trials, the subjects were not shown any study sequences and were not given any feedback, so they had no opportunity to learn more about the target pattern. In addition, they were supposed to generate an instance of the target pattern from a new prompt (T instead of M), so the transfer problem consisted of an entirely different set of letters. If the subjects acquired a concrete representation of the target letter sequence, with no representation of the underlying relational pattern, they would have found the transfer problems impossible to solve. This is equally true for all four feedback regimens. However, performance on the transfer problems was almost identical to performance on the assessment problems. The overall mean for the two assessment trials was 8.74 letters correct and for the two transfer trials was 8.44 letters. (This level of performance can be compared to the overall mean on the first two learning trials, which was 3.02 letters.) Clearly, the subjects could transfer what they had learned to the task of generating a instance beginning with an unfamiliar prompt. The conclusion is that they did not primarily use the letter-oriented strategy.

An interesting and important question is whether feedback regimen affected the balance between the pattern-oriented and letter-oriented strategies. If so, there should be differences between the groups in their ability to perform the transfer tasks, over and above the differences in assessment performance. But as Figure 4 and 5 show, not only was the mean performance on assessment and transfer similar, but the differences between the groups were numerically very similar in the transfer problems as in the assessment problems. The differences in transfer performance appear to be accounted for by the differences in the effectiveness of learning.

On the other hand, when we computed the difference between performance on the assessment and transfer trials, there did appear to be an effect of feedback. Figure 6 shows that the two groups that perform best after training show a performance drop when given a transfer task. This finding is consistent with the guidance hypothesis {Schmidt, 1989 #75}, suggesting that learner may become overly reliant on feedback as they learn a new skill.

Experiment 2

In Experiment 1, all subjects received both positive and negative feedback. Both types of feedback were delivered with the same probability, but this does not imply that the subjects necessarily received equal amounts of each type. The balance between positive and negative feedback is necessarily a function of the subjects' own behavior. If a subject makes many errors, there might be little opportunity to provide positive feedback, and vice versa. In our computer simulations, we have seen an interaction between scope and type of feedback (Ohlsson & Halpern, 1998), and the argument from information content implies that scope should matter more for negative than for positive feedback. The interaction between scope and probability that we observed in Experiment 1 might therefore be the result of an interaction between scope and type, superimposed on a difference in the actual amount of feedback received of either type.

In Experiment 2, we explore this hypothesis by separating the two types of feedback. Subjects received either positive or negative feedback, but not both, and either local or global feedback. The argument from information content predicts that there should be a main effect of feedback type in favor of positive feedback, because negative feedback provides less information and requires more interpretation. For negative feedback, there should also be an effect of scope in favor of local feedback, because global negative feedback requires more processing. The learner has to figure out which part or parts of a group of responses are wrong before he or she can figure out why they are wrong and what the correct responses are. However, neither the argument from information content nor the principle of cognitive load predicts any effect of scope on positive feedback. The information that a group of four responses is correct does not provide any different information, or require more processing, than four pieces of information that says that each of those four responses is correct.

Method

Participants. The subjects were 94 undergraduate students enrolled in the introductory psychology course at the University of Illinois at Chicago. They received course credit in return for their participation. The participants were 60% female and 40% male. No other demographic data were collected about the participants.

Materials. The training materials and the learning, assessment and transfer problems for Experiment 2 were identical to the corresponding materials for Experiment 1.

Design and Procedure. The subjects were randomly assigned to four groups, created by crossing type of feedback (positive or negative) with scope (local versus global). The subjects in the local scope groups received feedback that referred to individual letters, while the subjects in the global scope groups received feedback that referred to groups of four letters. Feedback probability was not varied. All subjects received 75% feedback of the relevant type. The procedure was the same as in Experiment 1.

Results

Verification of Hierarchical Task Representation. We computed response times for each position of the pattern. We analyzed the last four trials of learning, so that we could capture subjects' final representation of its structure. Figure 7 shows the result. The inter-response times are similar to those shown for experiment 1, providing additional support for the hypothesis that this pattern is represented hierarchically.

Learning. A Mixed repeated-measures ANOVA was performed to evaluate performance during the learning stage. All 12 learning trials were entered as a within-subjects factor. Feedback scope (local or global) and feedback type (positive or negative) were entered as between-subjects factor. The dependent measure was the number of letters correct per trial. There was a significant learning effect as shown in Figure 8, F (990, 11) = 42.21, p < .001. There was no effect of feedback scope, F (90 , 1) = 1.33, p > .05. There was no effect of feedback type, F (90 , 1) = 0.31, p > .05. There was an interaction between scope and type, F (90 , 1) = 8.45, p = .005. This interaction was due to the fact that subjects in the local negative group and the global positive group out performed the other two groups, F (90 , 1) = 8.38, p < .01. It was also reflected by the fact that subjects receiving negative feedback performed better when that feedback was given locally, F (90 , 1) = 8.25, p < .01. Subjects given positive feedback performed best when feedback was given globally, but this did not reach statistical significance, F (90 , 1) = 1.53, p > .05.

Assessment. A mixed within-subjects ANOVA was performed for the assessment trials. Both trials of the assessment stage were entered as within-subjects measure. Feedback type and feedback scope were entered as between-subjects measures. The dependent measure was the number of letters correct per trial. There was no effect of feedback type, F (90 , 1) = 0.10, p > .05 and no effect of feedback scope F (90 , 1) = 3.23, p > .05. The interaction between type and scope was significant, F (90 , 1) = 7.70, p < .01. This interaction was again due to the fact that subjects in the local negative group and the global positive group out performed the other two groups, F (90 , 1) = 7.75, p < .025. Figure 9 shows the results for the assessment trials. Subjects receiving negative feedback performed better when that feedback was given locally, F (90 , 1) = 10.45, p < .005. While subjects given positive feedback performed better when it was given globally, this trend did not reach statistical significance, F (90 , 1) = 0.48, p > .05.

Transfer. First, the two transfer trials were entered as a within-subjects measure in an ANOVA. Feedback type and scope were entered as between-subjects measures. The dependent measure was the number of letters correct per trial. Again, there was neither an effect of type nor of scope, F (90 , 1) = 0.05, F (90 , 0) = 2.15, p > .05. The interaction between type and scope was again significant, F (90 , 1) = 7.58, p < .01. Figure 10 shows the results for the transfer trials. Again, subjects in the local negative group and the global positive group out-performed the other two groups, F (90 , 1) = 7.61, p < .025. Subjects receiving negative performed better when that feedback was given locally, F (90 , 1) = 8.91, p < .005. Subjects given positive global feedback performed better than subjects given local positive feedback, but this did not reach statistical significance F (90 , 1) = 1.08, p > .05.

We also compared performance during assessment to performance during transfer. A comparison of figure 9 and figure 10 suggest that there is little difference between the two stages of learning. We entered stage (average performance on assessment verses average performance on transfer) as a repeated measure in an ANOVA. Feedback probability and scope were entered as between groups measures. There was no main effect of problem stage, suggesting that subjects performed equally well during assessment and transfer, F (100 , 1) = 1.038, $p > .05$. There was no interaction between scope, type and stage, F (100 , 1) = 0.08, $p > .05$.

Discussion

The results are consistent with some of our predictions, but in violation of others. There was no main effect of type of feedback. Contrary to the implications of the argument from information content and to the widespread belief in the power of positive feedback, positive feedback was not consistently superior to negative feedback.

The effects of scope were partially in accord with expectations. There was no main effect of scope and the expected interaction appeared, but not quite in the expected form. The expected difference between local and global feedback was significant for negative feedback. Negative feedback is more helpful when given locally and less so when given globally.

The results are inconsistent with the hypothesis that the effectiveness of feedback is solely a function of the amount of information it provides. If that were the case, then local negative feedback should not have been more effective than local positive feedback, and positive feedback should have been equally effective whether local or global. The results are consistent with the hypothesis that in a hierarchical representation, negative feedback that refers to higher level nodes is more difficult to process than negative feedback that refers to terminal nodes.

The differences between the groups observed during learning persisted during the assessment trials. The subjects performed very similarly on the transfer trials. The overall mean performance on the assessment trials was 9.4 letters correct and on the transfer trials 9.2 letters; these compare favorably with an overall mean of 3.8 on the first two learning trials. Clearly, the subjects learned something about the pattern that they could apply to the transfer tasks. However, once again, the differences between the groups on the transfer problems are highly similar to the group differences on assessment. There is no evidence that the different feedback regimens affected the subjects' ability to transfer. Unlike the previous experiment, there are no differences in learning gains from learning to transfer. The groups acquired different amounts of knowledge, but whatever the subjects in a group did learn, they could transfer.

General Discussion

From the point of view of common sense, the role and function of feedback in the acquisition of cognitive skills appear simple. When the learner receives positive feedback, he or she consolidates the correct step or steps in memory; when he or she receives negative feedback, the error is corrected. In general, the more feedback, the better. However, it might not be helpful merely to know that an action was in error, so the latter type of feedback might not be effective unless it also specifies the remedy in some detail.

Our findings show that this view of feedback is oversimplified to the point of being misleading, and contrary to fact in some respects. We found that neither the effects of feedback probability nor the effects of feedback type can be understood without taking into account the relation between the feedback and the organization of the learner's task knowledge. There is no average effect of either probability or type. Both variables interact with scope, i.e., whether the feedback refers to a single response or to a sequence of responses that constitute a meaningful component of the task solution. For local feedback, higher probability is indeed associated with faster learning; for global feedback, the opposite is the case. For local feedback, negative feedback leads to faster learning; for global feedback, the opposite is the case.

To understand these findings, several cognitive principles are useful. The argument from information content – that claim negative feedback is less effective than positive because it provides less information – cannot be accepted as it is usually stated (Hilgard & Bower, 1966), because local negative feedback turns out to be more effective than local positive feedback. Instead of focusing on the amount of information, we need to focus on amount of processing. Negative feedback requires complex processing; why is the action in error, and what is the remedy (Ohlsson, 1996b) In combination with the principle of hierarchical organization (strongly supported by the structure of our subjects' inter-response times; see Figures 2 and 7) and the idea that high cognitive load can interfere with learning (Sweller, 1994), the need for processing explains the effect of scope on negative feedback. The higher up in the hierarchical organization the negative feedback applies, the more complex is the processing required to correct the error. Hence, lower level (narrow scope) is preferable. For the same reason, increasing the probability of negative feedback only helps when scope is narrow. Increasing feedback probability for global negative feedback overwhelms the learner's capacity to process it and so interferes with learning.

Our data suggest that positive global feedback is more effective than local positive feedback. Although this trend did not reach statistical significance, there are two reasons to consider this trend important. First, this trend held in all three stages of learning. It was obtained during learning, during assessment and during transfer.

Second, positive global feedback is more efficient than local global feedback. Since positive global feedback is given only after a learner has completed all required responses correctly, it will come less often. Even if it were to turn out that positive local and positive global feedback were equally effective, there would still be good reason to chose the global feedback. In our study, a comparison of local to global positive feedback was done using the same probability (75%). This resulted on average 57 separate feedback message during learning in the local condition, as compared to 13 messages in the global condition. One could argue that each of these messages provides 4 separate messages. Multiplying 13 by 4, we find that on average subjects receive information concerning 42 letters in the global condition. It appears that global feedback achieves its effect more efficiently and using less information.

The argument from information content then fails to accurately describe the effect of scope on positive feedback. From an information content point of view, saying that the response sequence A, B, C, D is correct is equivalent to saying that response A is correct, response B is correct, and so on. Nevertheless, our data suggest that global positive feedback is more helpful than local.

To understand this effect, we need to turn the argument from information content on its head: To produce the correct response, the learner must already know the correct response. Hence, local positive feedback provides no information that the learner does not already have, and it is therefore less helpful than local negative feedback; the latter at least stimulates the search for a more correct alternative. Finally, to understand the effect of global positive feedback, we have to consider the effect of feedback on the learner's attention allocation. Feedback messages that refer to a component of the task reveal the components of the task and directs the learner's attention to those components. Positive feedback of higher than minimal scope carries two messages: it confirms correctness, but it also communicates task structure in way that local positive feedback does not.

The cognitive literature contains many attempts to formally model the acquisition of cognitive skills (Klahr et al., 1987). A minimal requirement for such models is that they contain learning mechanisms that can make use of both positive and negative feedback from the environment. Some theories, e.g., the current version of the ACT-R theory developed by John R. Anderson and associates (Anderson & Lebiere, 1998), appear to have no mechanism for translating either positive or negative feedback messages into an improved skill.

The present findings pose additional challenges for cognitive models of skill acquisition, over and above the minimal requirement of being able to learn from feedback. To be psychologically plausible, mechanisms for processing feedback must operate on hierarchical task representations and they must be affected by the scope of the feedback. Our prior work on modeling learning focused primarily on learning from local negative feedback (Ohlsson, 1993; Ohlsson, 1996b; Ohlsson & Rees, 1991) but did not consider either scope or probability, nor did we model the effects of positive feedback. We know of no formal model that can replicate the full set of results reported here.

The informal explanation for our results presented above has implications for educational practice. First, negative feedback should be supplied locally and at a high probability. Interestingly, this principle is already implemented in the line of successful intelligent tutoring systems that use the so-called model tracing technique (Anderson, Corbett, Koedinger, & Pelletier, 1995; Corbett, Anderson, & O'Brien, 1995). These systems follow the learner's problem solution step by step, and judge each one with respect to correctness. The student is given feedback as soon as he or she deviates from what the system has been programmed to consider the correct solution path. Evaluation in ecologically valid contexts have confirmed that tutors that operate this way are effective (Koedinger, 2001); see also Mitrovic and Ohlsson (1999) for a related result. Model tracing tutors are sometimes criticized because they do not allow the student to flounder and hence, the argument goes, do not teach the students how to recover from error. The obvious remedy would be to program the tutoring system to delay intervention and provide negative feedback with a more comprehensive scope, for example by providing a message such as "You need to consider your approach". However, our data strongly imply that negative global feedback should be provided sparingly, if at all.

Second, although educators tend to emphasize the beneficial effects of positive feedback, such feedback is not always helpful. Local positive feedback ("this step here is correct") provides the learner with almost no information. It is only helpful in confirming that a guess or a tentative response happens to be correct. Positive feedback should be designed primarily to focus the learner's attention on the meaningful components of the task. Applying this principle in practice requires a cognitive task analysis (Schraagen, Chipman, & Shalin, 2000) to identify those components. In most cases, those components will encompass more than a single elementary action.

The present studies have several limitations that might affect the generalizability of the findings. First, the subjects in our studies were under time pressure. They had only 45 seconds to study the feedback they received on any one attempt to extrapolate. The time pressure might have enhanced the effects of cognitive load and hence affected the effectiveness of global negative feedback. It is possible that subjects can learn more from global negative feedback than the present results indicate, if they are given unlimited time to process the feedback.

Second, the feedback messages were limited to "right" and "wrong". They did not provide any explanatory content ("this is wrong because ..."). Hence, these studies do not reveal what happens when such content is provided. It may seem as if richer information should help (McKendree, 1990), but empirical studies that compare content-poor and content-rich feedback regimens have not always found any differences (Corbett & Anderson, 1990). The principle of cognitive overload interfering with learning warns us to consider how much processing the extra content requires. The studies reported here shed no light on this issue, and it is possible that the effects or scope, probability and type are different in environments that provide content-rich feedback.

To advance our understanding of the role and function of feedback in the acquisition of complex cognitive skills, future studies need to address the unique methodological problems associated with this topic. Foremost among them is the fact that probability and type of feedback is a function of the subjects' behavior and hence not fully under experimenter control. A subject who makes few errors offers the experimenter few opportunities to issue negative feedback messages. The question arises how experimenters should define their feedback conditions. To retain fidelity to real learning scenarios, we choose to define our conditions in terms of the probability that feedback of a particular type would be received, given a response for which that type of feedback is appropriate. An experimenter might be tempted to use the power of computers to manipulate the actual amount of feedback received so that subjects in a given condition receive, for example, exactly 10 feedback messages of a given type. One consequence of this technique is to confound feedback probability with individual differences. The reduction of the number of positive feedback messages received, as compared to a 100% feedback condition, would obviously be larger for a subject with high cognitive ability than for one with a low cognitive ability; vice versa for negative feedback. The relation between behavior and feedback is intrinsically circular, and there is no right way to define feedback conditions. Researchers need to pay attention to how feedback conditions have been defined when comparing results from different studies.

A second methodological problem that has received too little attention in the past is the possibility that subjects interpret absence of feedback as indicating either correctness or error. This problem is particularly severe in conditions in which subjects receive only one type of feedback. We used an explicit "no feedback" signal to combat this problem, but we have no way of knowing how effective it was. There might be other techniques that work better in other task environments.

A third methodological issue is how we should measure the effects of different feedback regimens. Feedback might affect the process of learning, the representation of what is learned, or both. The former type of effect ought to show up primarily during practice, the latter primarily in what the learner can do with the acquired knowledge. We found no strong evidence for effects of the latter kind in the present experiments, but such effects might appear in other task domains or for skills of a different nature. The best approach is for researchers to capture as much data as possible in every study.

The role and function of feedback are essential parts of a theory of the nature and growth of cognitive skills. A theory of feedback is potentially of great practical importance. The recent neglect of the topic is presumably rooted in the mistaken belief that it has been researched extensively, that the basic effects and processes are well understood, and that the main truth about the matter is captured in the common belief that positive feedback is more helpful than negative feedback. The facts are otherwise. There is no systematic body of empirically grounded principles about the role and function of feedback in the acquisition of cognitive skills. Our experiments demonstrate that there are many effects and complex interactions that need to be better understood before such principles can be formulated.

References

Adams, J. A. (1971). A closed-loop theory of motor learning. Journal of Motor Behavior, 3(2), 111-150.

Anderson, J. R. (1976). Language, memory, and thought (Vol. xiii). Potomac, Md: Lawrence Erlbaum.

Anderson, J. R. (1981). Cognitive skills and their acquisition. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. (1982). Acquisition of cognitive skill. Psychological Review, 89(4), 369-406.

Anderson, J. R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1993). Rules of the mind: Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc. (1993). ix, 320pp.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. Journal of the Learning Sciences, 4(2), 67-207.

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

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. Psychological Review, 86(2), 124-140.

Arbib, M. A. (1964). Brains, machines, and mathematics. New York, NY: McGraw-Hill.

Ausubel, D. P. (1968). Educational psychology: A cognitive view. New York, NY: Holt, Rinehart and Winston.

Bilodeau, E. A. (1966). Acquisition of skill (Vol. xiii). NY: Academic Press. (1966).

Bourne, L. E. (1957). Effects of delay of information feedback and task complexity on the identification of concepts. Journal of Experimental Psychology, 54, 201-207.

Bourne, L. E., & Bunderson, C. V. (1963). Effects of delay of informative feedback and length of postfeedback interval on concept identification. Journal of Experimental Psychology, 65, 1-5.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8, 240-247.

Corbett, A. T., & Anderson, J. R. (1990). The effect of feedback control on learning to program with the Lisp tutor. In P. o. t. T. A. C. o. t. C. S. Society (Ed.) (pp. 796-803). Hillsdale, NJ: Erlbaum.

Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995). Student modeling in the ACT programming tutor. In P. D. Nichols & S. F. Chipman & e. al. (Eds.), Cognitively diagnostic assessment (pp. 19-41). Hillsdale, NJ, US; Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc; Lawrence Erlbaum Associates, Inc.

Corrigan-Halpern, A., & Ohlsson, S. (2001). Failure to Learn from Negative Feedback in a Hierarchical Adaptive System. In E. M. Altman & A. Cleermans & C. D. Schunn & W. D. Gray (Eds.), Fourth International Conference on Computer Modling. Fairfax, VA: Laurence Erlbaum.

Craik, K. J. W. (1947). Theory of the human operator in control systems. I. The operator as an engineering system. British Journal of Psychology (General Section), 38(2), 56-61.

Craik, K. J. W. (1948). Theory of the human operator in control systems. II. The opeator as an engineering system. British Journal of Psychology (General Section), 38(3), 142-148.

Egan, D. E., & Greeno, J. G. (1974). Theory of rule induction: Knowledge acquired in concept learning, serial pattern learning, and problem solving. In I. L. W. Gregg (Ed.), Knowledge and cognition (pp. 43-103). Potomac, MD: Wiley.

Greeno, J. G. (1974). Hobbits and orcs: Acquisition of a sequential concept. Cognitive Science, 6, 270-292.

Greeno, J. G., & Simon, H. A. (1974). Processes for sequence production. Psychological Review, 81(3), 187-198.

Hilgard, E. R., & Bower, G. H. (1966). Theories of learning (3rd ed. Vol. vii): East Norwalk, CT, US: Appleton-Century-Crofts. (1966).

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

Koedinger, K. R. (2001). Cognitive tutors as modeling tools and instructional models. In K. D. Forbus & P. J. Feltovich (Eds.), Smart machines in education: The coming revolution in educational technology (pp. 145-167). Cambridge, MA, US: The MIT Press.

Kotovsky, K., & Simon, H. A. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. Cognitive Psychology, 4(3), 399-424.

Kulik, J. A., & Kulik, C.-l. C. (1988). Timing of feedback and verbal learning. Review of Educational Research., 58(1), 79-97.

Langley, P. (1987). A general theory of discrimination learning, Klahr, David (Ed); Langley, Pat (Ed); et al. (1987). Production system models of learning and development. (pp. 99-161). xii, 466pp.

McKendree, J. (1990). Effective feedback content for tutoring complex skills. Human Computer Interaction, 5(4), 381-413.

Mesch, D. J., Farh, J.-L., & Podsakoff, P. M. (1994). Effects of feedback sign on group goal setting, strategies, and performance. Group & Organization Management, 19(3), 309-333.

Miller, G. A. (1998). Nouns in WordNet. In C. Fellbaum (Ed.), WordNet: An electronic lexical database (pp. 23-46). Cambridge, MA: MIT Pres.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). Plans and the structure of behavior. New York: Henry Holt and Company.

Mitrovic, A., & Ohlsson, S. (1999). Evaluation of a constraint-based tutor for a data-base language. International Journal of Artificial Intelligence and Education, 10, 238-256.

Moray, N. (1987). Feedback and the control of skilled behaviour. In D. H. Holding (Ed.), Human skills (pp. 15-39). Chichester, UK: John Wiley.

Murray, T. (1998). Authoring knowledge-based tutors: Tools for content, instructional strategy, student model and interface design. Journal of the Learning Sciences, 7(1), 5-64.

Nadel, L. (1991). The hippocampus and space revisited. Hippocampus, 1(3), 221-229.

Newell, A. (1990). Unified theories of cognition: Cambridge, MA, US: Harvard University Press. (1990). xvii, 549pp.

Nokes, T., & ohlsson, S. (2000). An inquiry into the function of implicit knowledge and its role in problem solving. In L. R. Gleitman & A. K. Joshi (Eds.), Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society (pp. 829-834). Mahwah, NJ: Erlbaum.

Ohlsson, S. (1987a). Transfer of training in procedural learning: A matter of conjectures and refutations? In L. Bolc (Ed.), Computational models of learning (pp. (pp. 55-88)). Berlin, Germany: Springer-Verlag.

Ohlsson, S. (1987b). Truth versus appropriateness: Relating declarative to procedural knowledge. In D. Klahr & P. Langley & e. al. (Eds.), Production system models of learning and development (pp. 267-327). Cambridge, MA, US: The MIT Press.

Ohlsson, S. (1993). Abstract schemas. Educational Psychologist, 28(1), 51-66.

Ohlsson, S. (1996a). Learning from error and the design of task envirnonments. International Journal of Educational Research, 25(5), 419-448.

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

Ohlsson, S., Ernst, A. M., & Rees, E. (1992). The cognitive complexity of learning and doing arithmetic. Journal for Research in Mathematics Education, 23(5), 441-467.

Ohlsson, S., & Halpern, A. (1998). Strength adjustment in hierarchical learning. Paper presented at the 20th Annual Conference of the Cognitive Science Society, Madison, WI.

Ohlsson, S., & Jewett, J. J. (1997). Ideal adaptive agents and the learning curve. In J. Brzezinski & B. Krause & T. Maruszewski (Eds.), Idealization VIII: Modeling in psychology. Amsterdam, The Netherlands: Rodopi.

Ohlsson, S., & Rees, E. (1991). The function of conceptual understanding in the learning of arithmetic procedures. Cognition and Instruction, 8(2), 103-179.

Patrick, J. (1992). Training: Research and practice. Londong, UK: Academic Press.

Proctor, R. W., & Dutta, A. (1995). Skill acquisition and human performance. Thousand Oaks, CA: SAGE.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. Psychological Review, 77(6), 481-495.

Restle, F., & Brown, E. (1970). Serial pattern learning: Pretraining of runs and trills. Psychonomic Science, 19(6), 321-322.

Ruiz, D. (1986). Learning and problem solving: What is learned while solving Tower of Hanoi? Unpublished Doctoral dissertation, Stanford University.

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: a review and critical reappraisal. Psychological Bulletin, 95(3), 355-386.

Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: support for the guidance hypothesis. Journal of Experimental Psychology: Learning Memory and Cognition, 13(2), 352-359.

Schraagen, J. M., Chipman, S. F., & Shalin, V. L. (2000). Cognitive task analysis. Mahwah, NJ: Erlbaum.

Schroth, M. L. (1997). The effects of different training conditions on transfer in concept formation. Journal of General Psychology, 124(2), 157-165.

Simon, H. (1972). Complexity and the representation of patterned sequences of symbols. Psychological Review, 79(5), 369-382.

Simon, H. A. (1962). The architecture of complexity. Proceedings of the American Philosophical Society, 106, 467-482.

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. Psychological Review, 70(6), 534-546.

Simon, H. A., & Sumner, R. K. (1968). Pattern in Music. In B. Kleinmuntz (Ed.), Formal representation of human judgement (pp. 219-250). New Yrok, NY: Wiley.

Skinner, B. F. (1938). The behavior of organisms. New York, NY: Appleton-Century-Crofts.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. Learning & Instruction, 4(4), 295-312.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. Cognition and Instruction, 12, 185-233.

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. Psychological Bulletin, 110(1), 67-85.

Thorndike, E., L. (1931). Human Learning. New York: Century.

Thorndike, E. L. (1913a). Educational psychology. Volume 1. The original nature of man. Teachers College, Columbia University.

Thorndike, E. L. (1913b). Educational psychology. Volume 2. The psychology of learning. New York, NY: Teachers College, Columbia University.

Thorndike, E. L. (1927). The Law of effect. American Journal of Psychology, 39, 212-222.

Thorndike, E. L. (1932). The fundamentals of learning. New York, NY: Teachers College, Columbia University.

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. Chicago: University of Chicago Press.

Welford, A. T. (1968). Fundamentals of skill. London, UK: Methuen.

Wiener, N. (1948). Cybernetics. New York, NY: Wiley.

Table Captions

Table 1: Twelve Exemplar Sequences that Instantiate the Same Pattern

Table 1.

Twelve Exemplar Sequences that Instantiate the Same Pattern

C A D F F D A C D B E G G E B D
E C F H H F C E F D G I I G D F
G E H J J H E G H F I K K I F H
I G J L L J G I J H K M M K H J
O M P R R P M O P N Q S S Q N P
Q O R T T R O Q R P S U U S P R
V T W Y Y W T V W U X Z Z X U W
K I L N N L I K L J M O O M J L
D B E G G E B D E C F H H F C E
F D G I I G D F G E H J J H E G
H F I K K I F H I G J L L J G I
J H K M M K H J K I L N N L I K

Figure Captions

Figure 1: Accuracy during training for all subjects for problems 1 and 2.

Figure 2: Inter-response times for the last four learning trials of Experiment 1.

Figure 3: Performance on 12 learning trials of Experiment 1.

Figure 4: Accuracy on assessment trials of Experiment 1.

Figure 5: Accuracy on transfer trials of Experiment 1.

Figure 6: Difference between assessment and transfer trials in Experiment 1.

Figure 7: Inter-response times for last four learning trials in Experiment 2.

Figure 8: Performance on 12 learning trials of Experiment 2.

Figure 9: Performance on assessment trials of Experiment 2.

Figure 10: Performance on transfer trials of Experiment 2.

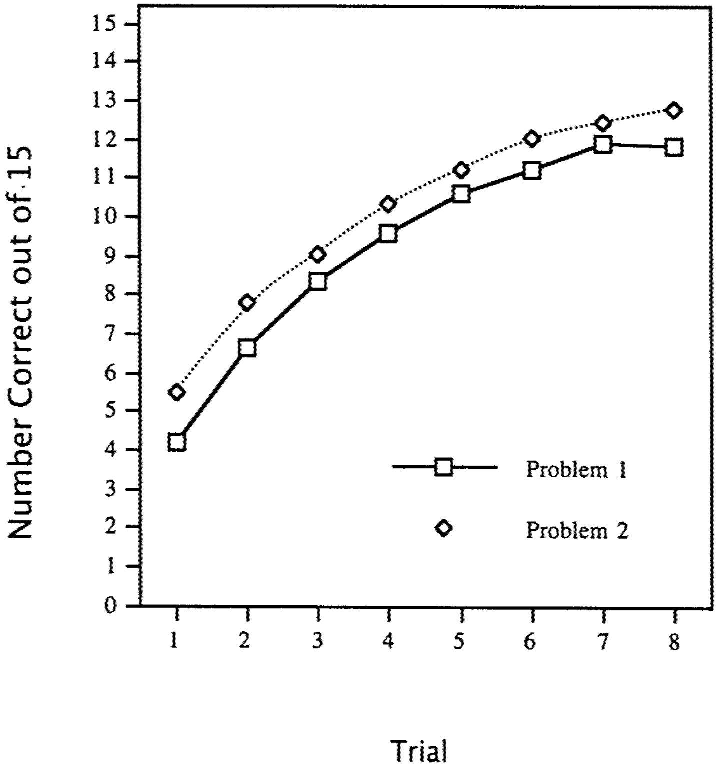


Figure 1. Accuracy during Training for All Subjects for Problem 1 and Problem 2.

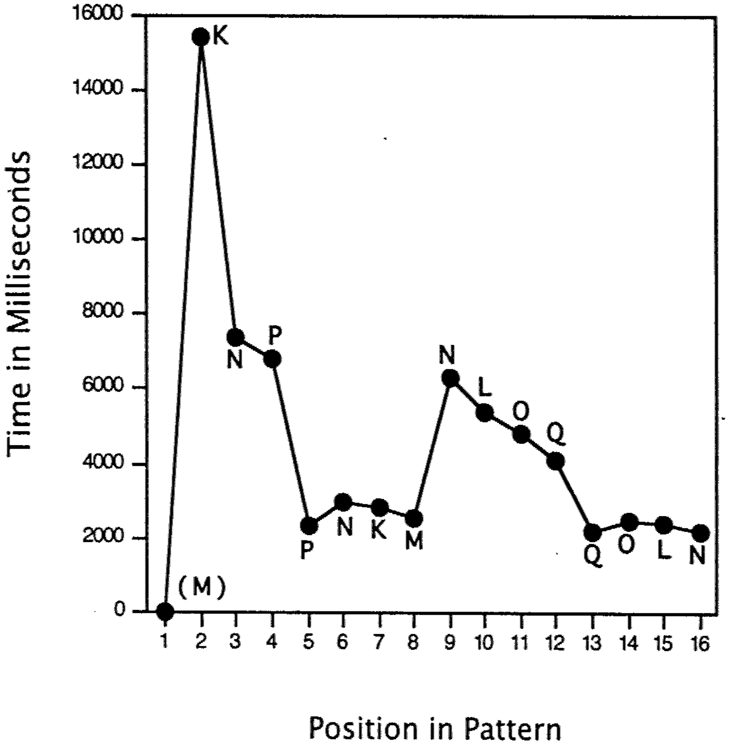


Figure 2. Inter-response Times for Last Four Learning Trials of Experiment 1.

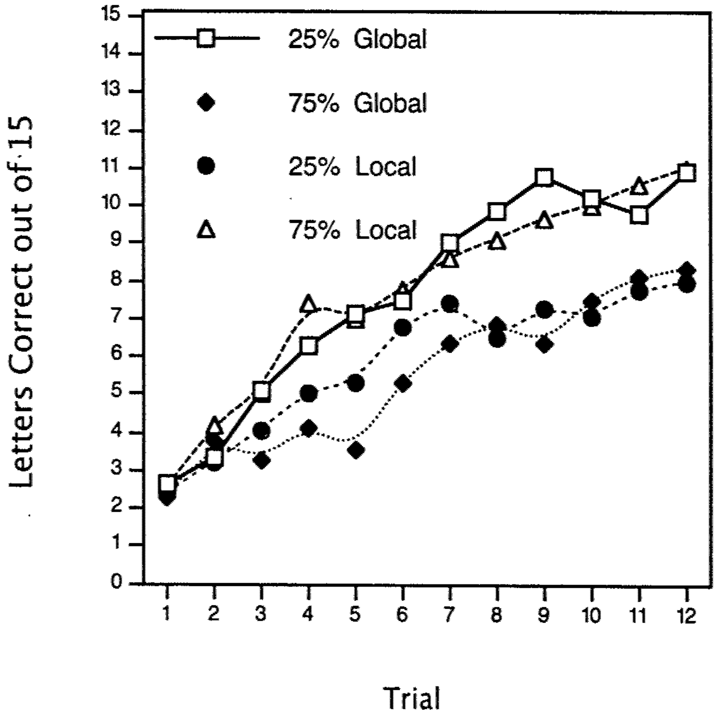


Figure 3. Performance for 12 Learning Trials of Experiment 1.

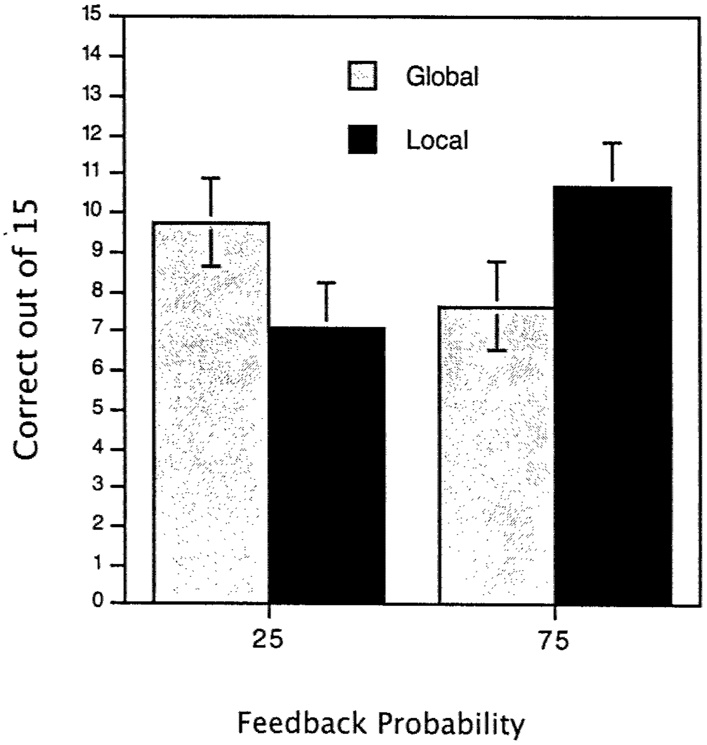


Figure 4. Accuracy for Assessment Trials of Experiment 1 with Standard Error.

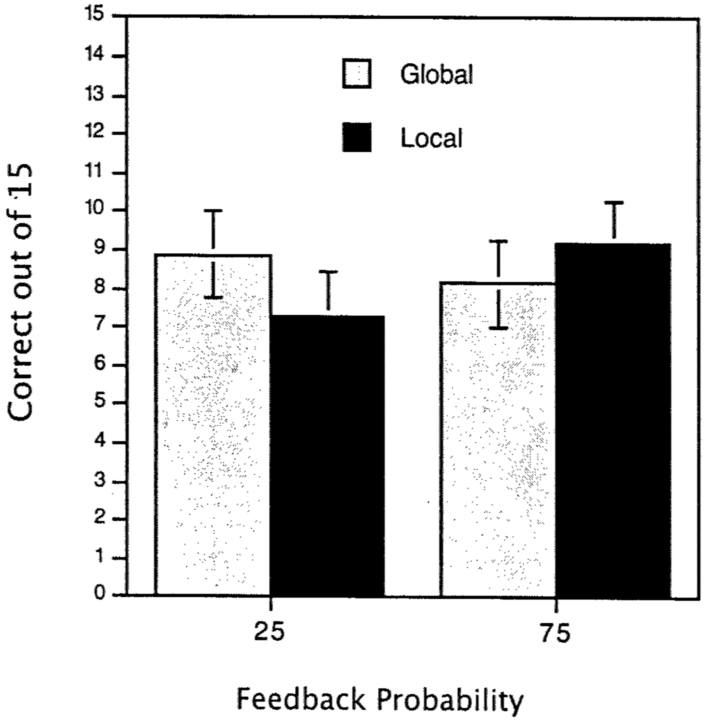


Figure 5. Accuracy for Transfer Trials of Experiment 1 with Standard Error.

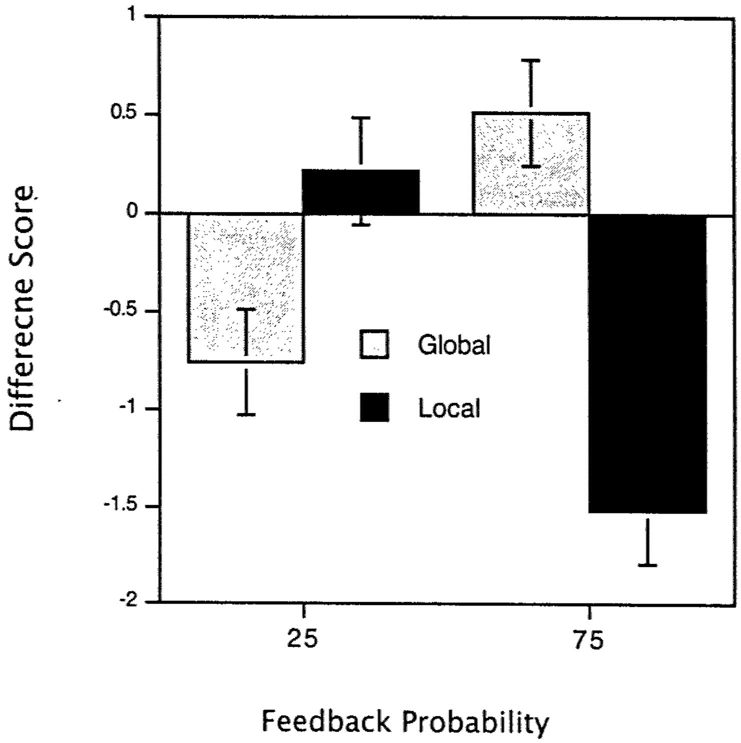


Figure 6. Difference between Transfer and Assesment Trials of Experiment 1 with Standard Error.

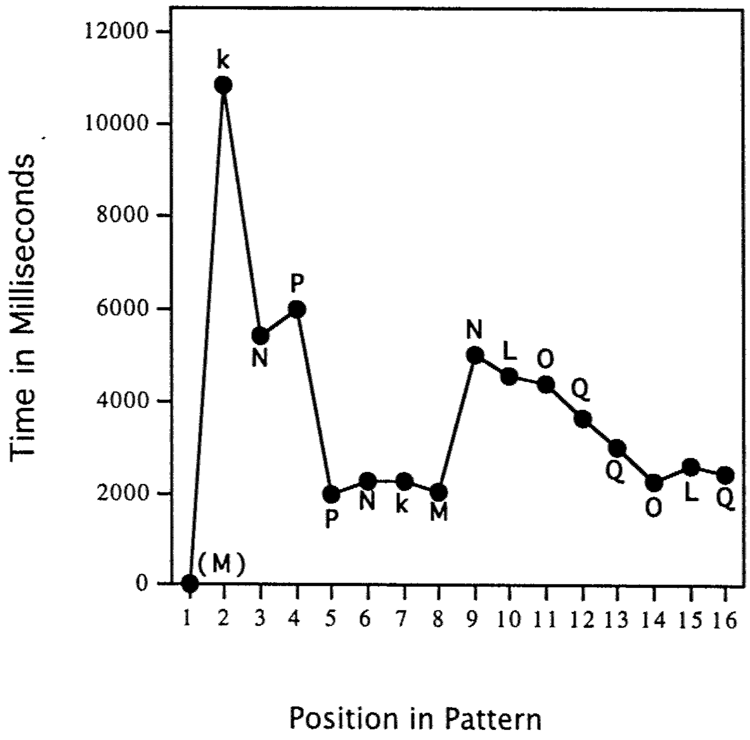


Figure 7. Inter-response Times for Last Four Learning Trials of Experiment 2.

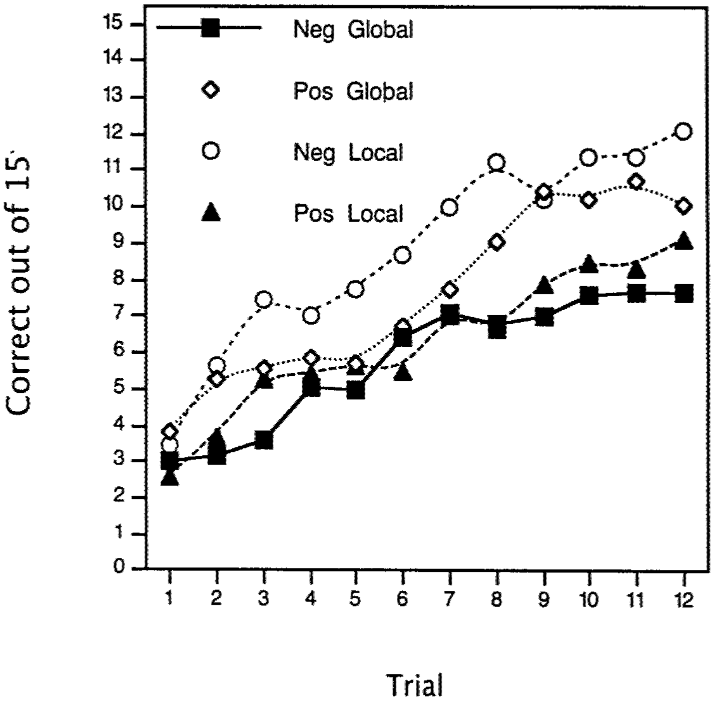


Figure 8. Performance for 12 Learning Trials of Experiment 2.

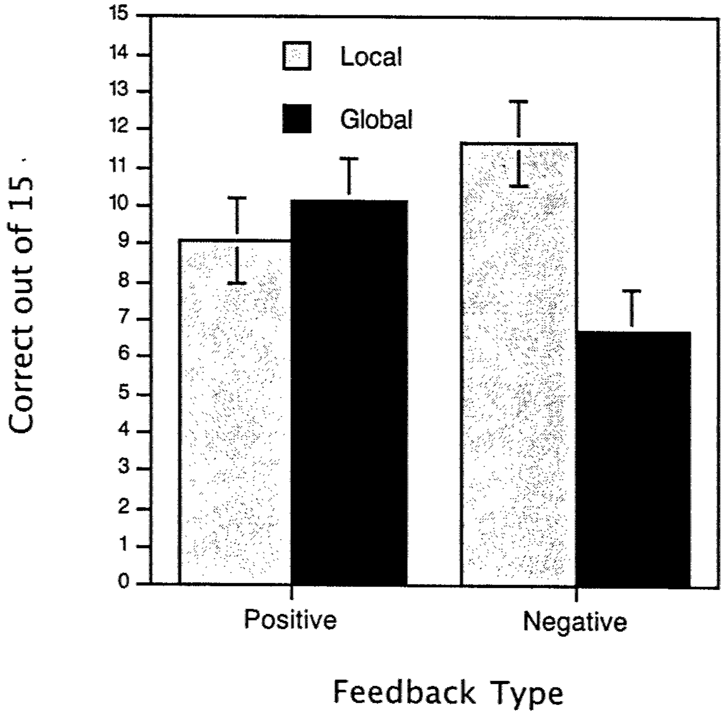


Figure 9. Performance for Assessment Trials of Experiment 2 with Standard Error.

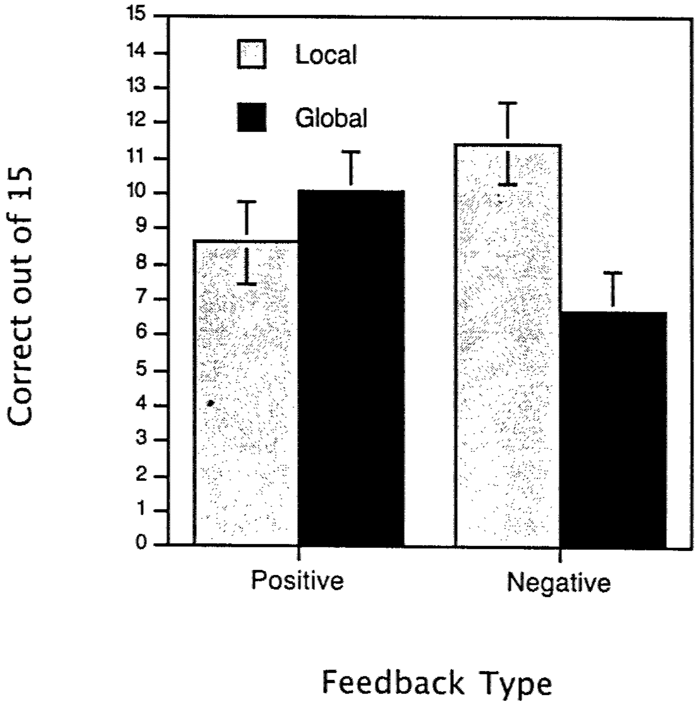


Figure 10. Performance for Transfer Trials of Experiment 2 with Standard Error.

Part II:

Multiple Paths to Mastery: What is Learned and Transferred?

BACKGROUND

In task domains like algebra, navigation and medical diagnosis, the purpose of learning is to be able to solve future problems in those domains. The impact of learning on subsequent problem solving is mediated by *knowledge structures*. Learning processes create or alter knowledge structures and those structures are later activated, applied and articulated vis-à-vis a current situation to generate behavior. The question of what is learned in a particular task environment is to be answered by specifying the relevant knowledge structures.

This seemingly straightforward view of the relation between learning and subsequent problem solving grows more complex when we consider the fact that people can master a task in different ways. For example, a person can learn how to assemble a lawn chair by studying written instructions that come with the chair; by being tutored by someone who already knows how to do it; by observing someone else do it; by practicing on other, similar but simpler contraptions; or by attacking the task via trial and error across multiple trials. Each of these *learning scenarios* provides different information to the learner and if a piece of information is presented in two of these scenarios, it is typically presented in different ways. The scenarios constitute alternative paths to mastery and their differences cannot be reduced to parametric variations (number of training trials, amount of feedback, etc.).

How does the principle that knowledge structures mediate between learning and problem solving account for the existence of multiple paths to mastery? One might argue that successful performance on a given target task requires a well-defined set of knowledge structures. All learning scenarios that produce successful performance on that task must therefore prompt the learner to create precisely those structures. According to this view, all effective scenarios lead to the same final state of knowledge.

This view is implausible for two reasons. First, there is no a priori reason to expect information received along different channels to be represented in one and the same way. For example, why should what is learned about the layout of a city by walking around become represented in memory in the same way as what can be learned by studying a map? Why should we expect a learner who is tutored in a task to acquire the same task knowledge as someone who attacks the task via trial and error, and why would the knowledge be represented in the same way in the two cases?

Second, a given behavior – a task solution – can be generated from multiple knowledge representations. For example, a person can decide how to walk from place A to place B in an unfamiliar city either on the basis of a memorized route (*walk X Avenue north for two blocks, turn left*, etc.) or on the basis of a mental map. Different cognitive processes are needed to generate the target behavior from these two representations, but the behavior itself, the act of walking, is the same. The fact that two learners behave in the same way does not imply that their behavior is based on the same knowledge structures.

We propose instead that knowledge structures created in qualitatively different learning scenarios differ with respect to which information they contain and how that information is represented. Different scenarios can lead to the same level of mastery, because correct, effective or successful behavior vis-à-vīs a task can be generated from those different knowledge structures. According to this view, it is reasonable to ask: How do the knowledge structures created in scenario X differ from those created in scenario Y? A general theory of human learning should be able to predict, for a given learning scenario, which knowledge structures will be created, and, for any pair of such scenarios, how the resulting knowledge structures will differ.

Contemporary work on learning follows a research strategy that does little to move the field towards this goal. A typical cognitive learning theory describes a single learning mechanism (e.g., analogy). To provide empirical support for its psychological reality, the researcher studies a learning scenario for which that mechanism has a high degree of face validity (e.g., to verify that people learn via analogy, give subjects analogues to a target problem and verify that performance improves). Further support for the hypothesized learning mechanism is accumulated by showing that it can explain the behavioral effects of various parametric variations of that scenario (e.g., number of prior analogues). However, unless the theoretician is willing to claim that his or her proposed learning mechanism is responsible for all learning (a claim that in most cases could be quickly and decisively falsified), the question how people learn in situations in which the mechanism is not applicable or plausible is left unanswered (e. g., how do people learn when they have no useful analogue in memory?).

A useful complementary research strategy is to compare qualitatively different learning scenarios with respect to what is learned. In educational research, alternative learning scenarios are often compared with respect to their effectiveness. That is, with respect to how well students perform after spending comparable amounts of time in two different scenarios. This type of study addresses different questions than those that are most germane to a cognitive psychologist. We are interested in the differences between the knowledge structures acquired by learners who perform at the same level but who arrived at that performance level along different routes. Empirically documented differences between the knowledge structures acquired in different but effective scenarios constitute phenomena that a general learning theory ought to be able to explain.

There are two broad classes of learning scenarios which, taken together, account for a significant proportion of human learning in both formal and informal contexts. One class contains situations in which the learner is presented with oral or written discourse that explicitly expresses the target knowledge that he or she is supposed to acquire. Lectures and textbooks exemplify this type of instruction. The learning that goes on in this type of scenario is sometimes described as *learning by being told.* We will refer to this type of scenario as *direct instruction*, or just *instruction* when the context prevents ambiguity. The key defining characteristic of direct instruction is that the instructing agent – person or machine – communicates the target knowledge in explicit form, usually via discourse.

A second important class of learning scenarios contains situations in which the learner engages in some activity that approximates the desired target performance. The learning that goes on in this type of scenario is usually called *learning by doing.* Scenarios of this type differ with respect to the variability of the practice problems. When there is minimal variability so that the learner is asked to perform the same task over and over again, it is often referred to as *drill.* When the variability is so great that the task of figuring out the connections between one practice task and the next is itself an intellectual challenge, it is called *analogical learning.* The term *practice* is typically used to refer to series of problems that are intermediate in variability between these two extremes. The exercises at the end of a textbook chapter in algebra is one example. The common feature of this class of learning scenarios is that the instruction is indirect; the target knowledge is not communicated explicitly. Instead, the learning activities – the practice tasks – are designed in such a way that by attempting to perform them, the learner is prompted to construct the target knowledge. In this article, we restrict the term "practice" to situations in which the learner is aware of the target performance that he or she is trying to achieve.

Learning by being told and learning by doing contrast with a third type of learning scenario that is variously called *incidental learning* and *implicit learning* (Postman, 1964; Reber, 1967, 1989; Richardson-Klavehn & Bjork, 1988; Seger, 1994; Schacter, 1987; Wattenmaker, 1999). In this type of scenario, the relation between the activities that the learner is asked to perform and the target task is even more indirect than in the case of learning by doing. The learner is not told what the target task is; indeed, he or she is not told that there is a target task. In educational contexts, instruction in arithmetic for young children with so-called manipulates exemplify this type of learning scenario. In this approach, instructors fashion material objects (blocks, bundles with sticks, etc.) in such a way that they follow the rules of the number system. They then invite children to engage in various activities with respect to these objects, e.g., trade them. The hope is that the children will extract important concepts about quantity and number in the course of these activities (Dienes, 1963; Skemp, 1971; but see Friedman, 1978, and Sowell, 1989, for some negative evidence).

The training materials used in laboratory studies of incidental learning tend to be very different form those used in educational contexts. For example, in the training phase of the standard artificial grammar learning scenario (Reber, 1967, 1989, 1993), the participants memorize symbol sequences (usually letter sequences). The sequences have been generated by an artificial grammar and hence share some properties, but the participants are not informed of this fact. In the test phase, the participants encounter new symbol sequences that are also derivable from the relevant grammar, mixed with distractors that are not. Their task is to decide whether the test sequences are of the same type as – i.e., derived from the same grammar as – the strings seen during the training phase. A large body of evidence (Berry, 1997; Reber, 1993; Stadler & Frensch, 1998) shows that people perform better than chance on this task. The key features of incidental learning scenarios as we use this term in this article are that the learning task does not approximate the target task and that the learner does not know at the time of learning what the target task is, or even that there is a target task.

These three classes of scenarios are obviously different and the differences cannot be reduced to parametric variations. If three learners master one and the same task in these three ways, it is plausible that they will acquire different knowledge structures. In what terms are those differences to be described? Cognitive science provides a vocabulary of useful concepts for discussing knowledge structures (Markman, 2000). We focus on two dimensions that are central to understanding the relation between learning and problem solving.

Dimensions of Knowledge Structures

Declarative versus procedural knowledge. Declarative knowledge is knowledge about the way the world is; it is descriptive in character and it can be evaluated with respect to veridicality (Ohlsson, 1994). Common facts such as *the winter is cold in Minnesota* are prototypical examples. Declarative knowledge is task-independent, e.g., it is not encoded in memory in the context of particular goals or actions. When applied to a task, it has to be interpreted with respect to its consequences for action. For example, the declarative principle that three congruent sides makes two triangles congruent implies that two triangles can be proven congruent by proving that their sides are congruent (Neves & Anderson, 1981). The process of deriving the action consequences of declarative knowledge is called *proceduralization* or *knowledge compilation* (Anderson, 1983; Ohlsson, 1996).

Procedural knowledge is knowledge about how to perform tasks and, more generally, about how to achieve particular types of results in certain types of situations (Ohlsson, 1994). Everyday skills like driving and cooking are prototypical examples. Procedural knowledge is task-specific, so its application is quick and efficient. It can be evaluated with respect to appropriateness and effectiveness.

Although artificial intelligence researchers abandoned the declarative-procedural distinction in the 70s (Winograd, 1975), the distinction has turned out to be useful for psychology. Not only is it easy to point to intuitively clear examples of declarative and procedural knowledge – maps versus routes, theorems versus proof procedures – but there is also strong support from neuropsychology (Squire, 1987). People with particular types of brain damage can learn new skills – acquire new procedural knowledge – even though they might not be able to remember everyday events or acquire other types of declarative information.

John R. Anderson has argued for the centrality of this distinction in human cognition:

> *"It seems that certain knowledge can be best represented declaratively and other knowledge can be best represented procedurally. It is much more economical to represent declaratively that knowledge which is subject to multiple, different uses and that knowledge whose [sic] eventual use is uncertain. ... On the other hand, knowledge that is used over and over again in the same way ... would seem to be better represented in a procedural format in which it can be applied more rapidly. ... procedural knowledge is specific to the circumstances where it is intended to apply ... ."* (Anderson, 1976, p. 118)

A person's knowledge about a task can be either declarative, procedural or some mixture of the two. The two types of knowledge trade off applicability and efficiency in opposite ways: Declarative knowledge is widely applicable but to derive its consequences for action is a slow and complex process. Procedural knowledge applies only in a narrow task context but the application is fast and efficient. For any learning scenario we can ask whether it generates declarative or procedural knowledge, or some mixture of the two.

Abstract versus specific. A second key property of knowledge structures is their level of abstraction. The two terms "abstraction" and "generality" are used interchangeably both in ordinary speech and in cognitive discourse. However, we suggest that the informal terms cover two distinct concepts. In our view, the *generality* of a representation is a function of the number of instances in the world that match that representation. The concept *dog* has more instances than the concept *poodle*, so the former is more general. Generality admits of degrees, because a representation can have 0, 1, 2, … , N instances.

In contrast, *abstraction* is not a relation between a knowledge structure and the world but a property of the knowledge structure itself. A mental representation is abstract to the extent that it encodes relational information but leaves out information about the identity and the specific attributes of objects and events (Ohlsson, 1993; Ohlsson & Lehtinen, 1997). The opposite of *abstract* is *specific*.

The distinction between abstraction and generality is necessary because it is possible for a knowledge structure to be abstract and yet not have any instances. The concept *perfect justice* might qualify as an example. Mathematics and science provide many others. There are mathematical theories that have found no application to the physical world; they have no instances at all and their generality is therefore minimal. This does not make those theories less abstract than those mathematical theories that have found multiple applications, such as calculus. On the other side of the coin, two representations need not have different levels of abstraction in order to differ in generality. For example, there a good many more beetles in the world than there are mammals, but we do not think of the concept *beetle* as more abstract than *mammal*. Some fully specified concepts have no generality (e.g., unicorn) while some abstract concepts have many instances (e.g., non-linear functional relation).

Abstraction is of central concern in studies of learning, because people are able to transfer what they learn in one context to other contexts. However, in systematic studies of learning, such transfer does not always come easily to the participants (Cormier & Hagman, 1987; Detterman & Sternberg, 1993; McKeough, Lupart & Marini, 1995; Perkins & Salomon, 1989; Salomon & Perkins, 1989; Royer, 1979). It is thus of particular interest to study under which circumstances abstract knowledge representations emerge.

Abstraction is not continuously graded but we can distinguish between two levels (Ohlsson, 1993). We refer to a knowledge structure as being of *intermediate abstraction* when it leaves objects and events unspecified but encodes relations as specifically as the context allows. When the relations themselves are encoded with minimal semantic features, we say that the knowledge structure is of *higher-level abstraction*. For example, suppose that P stands for some person and O for some wooden object and that *P carved O*. This proposition is of intermediate abstraction because the person and the object are left unspecified but the relation, *carved*, is rather precisely specified. *P produced O* is more abstract, because the manner in which P produced O has been left unspecified. Similarly, the relation *related to* is more abstract than specific kinship relations like *cousin of*, *child of*, etc.; the relation *next element* (in some sequence) is more abstract than either *next letter* (in the alphabet) or *next number* (among the natural numbers); and so on. The question is whether particular learning scenarios differ in whether the knowledge structures they generate are specific, of intermediate abstraction, or of higher-level abstraction.

Application to Three Learning Scenarios

How do the knowledge structures created by direct instruction, practice, and incidental learning differ on the two dimensions of declarative versus procedural and abstract versus specific?

Type of knowledge. The obvious expectation is that learning by being told generates declarative knowledge. Without any opportunity to practice, the learners should not acquire procedural knowledge. Hence, we would expect the application of the knowledge to a target task to be slow because the learner has to proceduralize the knowledge on the run, as it were. An everyday example is to translate spatial relations extracted from a city map into a route for getting to a target location, a task that many find difficult.

A second expectation is that learning by practicing generates procedural knowledge. The application of this type of knowledge should be faster, because proceduralization has already been carried out during prior learning. However, working on a task is likely to also generate declarative knowledge about that task. For example, practice in using a camcorder generates the appropriate procedures for inserting and removing the tape, starting and stopping the recording function, and so on, but it also presents the learner with declarative information about the camera, e.g., it is lightweight, it is silent, and so on. Practice is likely to generate both declarative and procedural knowledge.

Finally, which type of knowledge is constructed in incidental learning scenarios? To the extent that learners in this type of learning scenario are engaged in a particular type of activity, we must suppose that they acquire procedural knowledge relevant to that activity. However, the intent of those who create such scenarios, either in the classroom or in the laboratory, is that the learners unintentionally acquire knowledge about the task materials. Indeed, the operational definition of incidental learning is that engaging in activity A creates knowledge that, without intent on the part of the learner, affects performance on the designated but unmentioned target task B. In most instances, this transfer effect could only plausibly be mediated by declarative knowledge. From this point of view, we should expect the knowledge structures created by direct instruction and by implicit learning to be similar.

Level of abstraction. Although abstraction has been a topic of inquiry since antiquity, we do not know under which conditions individuals create knowledge structures at any particular level of abstraction. An additional weakness in prior research for present purposes is that discussions of abstraction have not consistently distinguished between declarative and procedural knowledge. However, the three levels of abstraction distinguished above – specific, intermediate, higher-level – apply to both types of knowledge. Assertions and descriptions can certainly vary in abstraction level, but so can procedures. Hence, the question of abstraction must be considered separately for the two types of knowledge.

Beginning with direct instruction, what abstraction level would we expect for the resulting declarative knowledge? In a previous study, we found evidence that our participants encoded declarative task knowledge at the intermediate but not at the higher level of abstraction (Nokes & Ohlsson, 2000). That is, they abstracted over particular objects, but we found no evidence that they spontaneously encoded relations at any level of abstraction higher than the one in which the task materials were presented. Although far from decisive, that study leads us to predict that learning scenarios that produce more declarative than procedural knowledge – i.e., direct instruction scenarios – will lead to knowledge structures of intermediate abstraction. The implication is that they should transfer their declarative knowledge easily to a target task that differs in specific details but share the same relational structure, but not so easily to a task that also differs in relational structure.

For practice scenarios, our expectations are different. In prior work on learning by doing (Ohlsson, 1996; Ohlsson, Ernst, & Rees, 1992; Ohlsson & Rees, 1991), we have proposed a *specialization principle*: Errors are unlearned during practice primarily by gradually constraining (specializing) the procedural knowledge until each component (production rule) becomes active only in those circumstances in which it leads to the correct action. Contrary to the widespread idea that knowledge is concrete at the outset and becomes abstract later, the specialization principle claims that procedural knowledge starts out as instances of abstract procedures, and becomes more and more specific as those initial procedures are adapted to the specifics of the particular task being practiced. The specialization principle thus predicts that indirect instruction should produce procedural knowledge that is specific rather than abstract. However, this expectation is moderated by the variability of the practice sequence: The more variable the sequence, the stronger the prompt towards an abstract representation (Salomon & Perkins, 1989).

With respect to incidental learning scenarios, there is a debate in the literature about the level of abstraction that is required to explain the relevant empirical findings. Servan-Schreiber and Anderson (1990) and Perruchet and Gallego (1997) have attempted to explain artificial grammar learning in terms of specific surface features of the training sequences. In particular, they propose that the participants in implicit learning experiments learn the relative frequencies of particular substrings (e.g., letter pairs). This type of statistical information about the surface features of the strings might be sufficient to bias the judgments during the test phase towards the right choices. In response, Manza and Reber (1997) reported a series of six experiments in which the sequences encountered during the test phase were expressed in different letters than the sequences encountered in the training phase. In this situation, knowledge about the relative frequencies of particular letter combinations does not help during the test phase. Nevertheless, people perform better than chance in this condition also. In addition, Altmann, Dienes, and Goode (1995) have shown that the acquired grammatical pattern transfers from sequences of tones to sequences of letters. In short, there is evidence that what is acquired in incidental learning scenarios is abstract, but the level of abstraction is under debate.

To turn these arguments into behavioral predictions, we need to apply them to a particular task environment.

TASK ENVIRONMENT

To investigate these issues empirically, we need to apply the arguments in the preceding section to a particular task. The class of suitable tasks includes algebra, card games and software use. However, ecologically valid tasks tend to be irregular, draw upon unspecifiable parts of the learner's prior knowledge, and be learned under conditions that hinder precise control and assessment of relevant variables. The ideal target task shares key properties of ecological cognitive skills but enables experimental control and data capture.

Our target task is a version of the *letter sequence extrapolation task* studied by Greeno and Simon (1974), Klahr and Wallace (1970), Restle (1970), Restle and Brown (1970), Simon (1972), Simon and Kotovsky (1963), and others. The learner is given a sequence of letters that exemplify a pattern and he or she is asked to continue the sequence in such a way that the continuation also fits that pattern. In the original version of this task (Thurstone & Thurstone, 1941), the participant was asked to add a single next element to the sequence. In the version we use, the participant is asked to extrapolate the sequence to N places. For example, to extrapolate the sequence MABMCDM to six places, the participant would produce EFMGHM. Although the solution to this example is obvious, sequence extrapolation problem can be quite difficult. The reader might want to extrapolate EFDGCOFGEHDP to 8 places.

To solve such a problem, the participant must first identify the pattern in the given sequence and then use that pattern to generate the continuation of the sequence. The first part, *pattern detection*, consists of studying the given part of the sequence to identify relations between the letters. The relations can be of various kinds: a letter can be identical to another letter, follow another letter in the alphabet, precede another letter, be two steps removed from another letter in the alphabet, and so on. Groups of letters can relate to each other in additional ways: A group of letters can be like another group in that the same relations hold between the letters (e.g., MCD is like MAB in this sense), a group of letters can be the reverse of another group (e.g., ABC and CBA), and so on. A pattern can have a periodic structure. In this type of pattern, the sequence is divided into *periods*, subsequences that are similar in the sense defined above. All the patterns used in the studies reported in this article are periodic. The pattern underlying a particular sequence can be said to have been understood when every letter in the sequence is seen as related to the other letters in some specified way. Understanding the pattern is a prerequisite for successful performance of the task.

The second part, *sequence extrapolation*, requires inferences, one for each position extrapolated. In periodic patterns, an extrapolation inference is based on an analogy between periods. For example, to decide that "E" is the correct next letter in the sequence MABMCDM, the problem solver has to realize that MAB and MCD are the two periods and that the first position to the right of the third M is analogous to the two positions to the right the first and second M. The latter two are occupied by A and C, which are related by the fact that C is two steps forward in the alphabet from A. Hence, the letter to the right of the third M should be two steps forward in the alphabet from C, which is E. In this representation of the pattern, the letter M serves as a *period marker*, and the identification of the periodicity of the pattern is crucial for correct extrapolation. Each position in the extrapolated portion of the sequence requires an inference of this sort.

The letter sequence extrapolation task has several properties that make it a suitable model of complex cognitive skills. (a) Prior knowledge. Like ecological tasks, the process of mastering a letter sequence extrapolation problem draws upon the participant's prior knowledge. However, that knowledge is relatively circumscribed, namely the alphabet. (b) Conceptual content. Detecting the pattern in the given sequence is potentially facilitated by knowledge of the pattern-forming operations that were used to generate the given sequences. (c) Types of knowledge. Knowledge of a pattern is a good example of declarative knowledge, while the skill of extrapolation is a good example of procedural knowledge. Both are equally important for successful performance in this task. (d) Levels of abstraction. The pattern in a particular sequence can be encoded without abstraction, in terms of the specific letters that appear in the sequence. It can also be encoded at an intermediate level of abstraction that ignores the specific letters but encodes the relational structure of the underlying pattern (Nokes & Ohlsson, 2000; Ohlsson, 1993). Finally, it can be encoded at a higher level of abstraction by noting which positions in the sequence are related, without specifying the relations. (e) Complexity and generativity. A sequence extrapolation problem cannot be solved by memorizing and recalling the given sequence, but requires that the participant generates a novel sequence of coordinated responses.

Most important for our current purpose, the correct extrapolation of a sequence is facilitated if the person has learned the underlying pattern ahead of time. Prior declarative knowledge of the pattern ought to facilitate pattern detection. That is, it should be easier to recognize a familiar pattern in a letter sequence than it is to identify an unfamiliar pattern. Similarly, practice in carrying out extrapolation inferences should generate procedural knowledge that makes future inferences easier. Both patterns and extrapolation skills should transfer to related problems if they are encoded at some level of abstraction. Sequence extrapolation tasks are thus well suited as instruments with which to study the learning outcomes produced by a variety of learning scenarios.

OVERVIEW OF STUDIES

The purpose of the three studies reported in this article was to compare what is learned when a letter sequence extrapolation task is mastered via three contrasting learning scenarios. Specifically, we compare learning by being told, learning by doing and incidental learning with respect to their effects on subsequent problem solving. If the knowledge structures acquired during the three different training procedures are the same, the participants' behavior on a target problem should be similar. On the other hand, if the knowledge structures differ with respect to either type of knowledge (declarative vs. procedural) or level of abstraction (specific, intermediate, or higher-order), then we expect to see differences in behavior.

Declarative knowledge requires proceduralization and hence should be slower in application than procedural knowledge. We therefore predict that direct instruction will produce longer solution times than practice, even for participants who perform at comparable levels of accuracy. On the other hand, the standard trade-off argument for the existence of two types of knowledge (Anderson, 1976, 1983; Winograd, 1975) implies that declarative knowledge, by virtue of not being indexed under particular goals or tasks, has a higher generality and transferability than task-specific procedural knowledge. Hence, we expect a relative advantage for declarative over procedural knowledge, and hence for direct instruction over practice, on a transfer task. Because variable practice is widely believed to generate more abstract knowledge than uniform drill, the relative advantage for instruction over practice should be greater when compared to low variability practice and lesser with respect to high variability practice.

Incidental learning should produce declarative but not procedural knowledge. The behavioral pattern should thus be similar to that of the direct instruction participants. However, this prediction assumes that participants in this scenario can access what they learned during the incidental training procedure and use it to support deliberate problem solving. In past work (Nokes & Ohlsson, 2000) we found a very limited ability to do so. Hence, we predict overall lower performance for the incidental training procedure on both target and transfer problems.

The plan of the rest of the article is as follows. We report three experiments that recorded participants' performance in four qualitatively different training scenarios: no prior training, direct instruction, practice, and incidental training. The latter three scenarios were implemented in two parametric variants each: direct instruction was either short or long; the practice sequence was of either low or high variability; and the incidental training sequence was either short or long. For each experiment, we report accuracy and solution time results. We also report the quantitative amount of transfer, using two of the transfer formulas reviewed by Singley and Anderson (1989). In a separate section, we analyze the structure of the solution times in detail in order to arrive at a processing account of our findings.

EXPERIMENT 1: BASELINE

The purpose of experiment 1 was to determine how participants solve the target problems in the absence of any prior training. In this *no-training condition* participants were given the letter sequence extrapolation in problem solving mode. That is, they were given the minimum of task instructions needed in order to be able to attack the problems, but no information regarding the nature of the particular patterns hidden in the letter sequences. This base line measure is needed to (a) establish that we do not have floor or ceiling effects that could artificially restrict the possible effects of training procedures explored in experiments 2 and 3, and (b) to serve as a comparison group when analyzing the data from experiments 2 and 3.

Method

Participants

Thirty undergraduate students from the University of Illinois at Chicago participated in return for course credit.

Materials

The target tasks were two letter sequence extrapolation problems. Problem 1 had a periodicity of 6 items and problem 2 a periodicity of 7 items; see Table 1. To enable the participants to detect the embedded pattern, the given segment was 12 items long for problem 1 and 14 items long for problem 2. That is, the given segments covered two complete periods of the patterns. These problems were created specifically for this experiment but are similar in character to the problems used by Kotovsky and Simon (1973).

Each target problem was associated with a transfer problem. The patterns embedded in the transfer problems were related, but not identical, to the patterns in the target problems. For problem 1, the corresponding transfer problem was generated by quantitatively 'stretching' the relations between the letters. For example, "forward 1 in the alphabet" was 'stretched' to "forward 2 steps", "backwards 1" was 'stretched' to "backwards 2", and so on; see Figure 1 for a detailed example. The second transfer problem was generated in the same way from target problem 2. This method of generating transfer problems preserves the qualitative structure of the target pattern (i.e., which positions are related and the direction of each relation) but changes the quantitative aspect of each relation.

Design and Procedure

All participants were assigned to a single condition.

They were tested in groups of 1-4 people. Each participant sat in front of a Macintosh computer with a 17' color monitor, standard keyboard and mouse. All stimuli were presented in black 30 pt font in the center of the screen; see Figure 2 for screen layout. The experiment was designed and presented using PsyScope software.

The procedure consisted of two *cycles*. Each cycle was composed of one target and one transfer problem. Participants were first given general instructions on how to solve sequence extrapolation problems. They were told that they would be given a sequence of letters containing a pattern and that their task would be to find the pattern and fill in the next 8 letters in such a way that the letters also followed the pattern.

Participants were given an example problem and solution; "ABMCDM continue the sequence . . . the next 8 letters of the solution are EFMGHMIJ". They were told that the given letters of a problem would be presented on the left side of the screen and that there would be an empty box for each of the 8 letters they were to extrapolate. To fill in a box they were told to click the mouse on the box they wanted to fill in after which a question mark would appear and they could press the appropriate letter; see Figure 2 for an example. The participants could fill in the positions in any order they choose. After all 8 letters were filled in participants were told to click the mouse on the "Finished" field.

After reading through the general instructions participants were presented the first target problem. Participants extrapolated each problem to eight positions; thus for problem 1 participants extrapolated one complete iteration (i.e., 6 positions) plus 2 positions of the next iteration. For problem 2 they extrapolated one complete iteration (i.e., 7 positions) plus 1 position of the next iteration. They were given 6 minutes to solve the problem. Next, they were given the transfer problem. Again they were given 6 minutes to solve the problem. The second cycle proceeded in the same way. Problem 1 and problem 2 were counter-balanced across all participants. The entire procedure took approximately 40 minutes.

Results

Because the task instructions directed the participants to extrapolate correctly but did not mention speed, accuracy is our basic performance measure. The *problem solving score* was determined by the number of letters correctly extrapolated in each problem solving task. Because participants were asked to extrapolate each problem to eight positions their problem solving scores varied from 0 to 8. Mean problem solving scores were 3.73 (*SD* = 3.32) on target 1, 3.20 (*SD* = 3.22) on transfer 1, 3.43 (*SD* = 2.81) on target 2, and 3.00 (*SD* = 3.02) on the transfer 2. Alpha was set to .05 for all subsequent tests.

Although Figure 3 shows slightly lower performance on the two transfer problems than on the two target problems, a one-way repeated measures analysis of variance (ANOVA) revealed a non-significant difference of problem (target 1 vs. transfer 1 vs. target 2 vs. transfer 2), *F* (3, 87) = 1.05, *MSE* = 2.82, *ns*, indicating that participants problem solving performance did not differ across problems. Participants correctly extrapolated an average of 3.34 positions.

Throughout we use the *solution time*, i.e., the total number of seconds the participant worked on each problem, as an indicator of amount of cognitive processing. Because participants were given a maximum of 360 seconds (6 minutes) to work on a given problem their time scores varied between 0 and 360. Mean solution times were

248.4 (*SD* = 85.8) for target problem 1, 253.2 *(SD* = 92.4) for transfer problem 1, 246 (*SD* = 99) for target problem 2, and 250.2 (*SD* = 95.4) for transfer problem 2.

Although Figure 4 shows that the participants solved the target problems slightly faster than the transfer problems, a one-way repeated measures ANOVA revealed a non-significant difference for solution times across problems (target 1 vs. transfer 1 vs. target 2 vs. transfer 2), *F* (3, 87) = .06, *MSE* = 1.50, *ns*. Solution time scores were similar to accuracy scores in that performance did not differ across problems.

Discussion

The two target problems turned out to be of approximately similar levels of difficulty. The participants correctly extrapolated less than half of the 8 positions and it took them more than four minutes to do so, so these are relatively difficult problems for our population of participants. However, they are not so difficult that the participants cannot make any progress. In short, we found neither floor nor ceiling effects. In addition, the fact that performance on the transfer problems was no better than on the target problems indicates that transfer of training is not straightforward in this task domain and that the transfer problems were not intrinsically more difficult than the target problems.

EXPERIMENT 2: PRACTICE VS. DIRECT INSTRUCTION

A large percentage of formal learning situations put the student in either a direct instruction scenario or in a practice scenario. On the assumption that direct instruction primarily prompts the construction of declarative knowledge, while practice primarily prompts the construction of procedural knowledge, these two scenarios, when applied to the same task environment, should generate different behaviors. The purpose of experiment 2 was to gather data that would allow us to evaluate this prediction. The general approach was to provide either instruction or practice before the participants encountered the target and transfer problems we explored in experiment 1. Performance on the target and transfer problems thus served as tools for assessing the effects of the prior training. We refer to the target and transfer problems collectively as the *test problems*. To what extent could the participants utilize either type of training to improve their performance on the test problems, as compared to the no-training group from experiment 1? If they could, were their task behaviors similar or different? If the latter, how did they differ?

Direct instruction consisted in this case of a written description of the pattern underlying the letter sequence in the target problem. The description used a graphical representation of the relational structure of the pattern (see below for further details). The learners studied the relations and recalled them. This scenario was implemented in two parametric variants, called *short instruction* and *long instruction*. The participants in the short variant studied a booklet that was 12 pages long. The participants in the long variant studied an additional 2 pages. Those two pages described the process of extrapolating a patterned sequence (see below).

Practice consisted in this case of three practice trials, with only minimal task instructions. On each trial, the participant solved a letter sequence extrapolation problem that embodied the same pattern as the target problem. This scenario was also implemented in two parametric variants. The participants in the *low-variability practice* variant solved exactly the same practice problem (same pattern, same letters) three times. The participants in the *high-variability practice* variant solved three practice problems that all used the same pattern, but instantiated it in different letters. The low-variability participants had no reason to create abstract knowledge before encountering the target task, but the high-variability participants did. The practice problems used the same pattern as the target problem, but there was no overlap with the target problem with respect to the letters used to instantiate the pattern.

Methods

Participants

One hundred and nineteen undergraduate students from the University of Illinois at Chicago participated in return for course credit.

Materials

The target and transfer problems were the same as in experiment 1; see Table 1. In addition, there were a total of three practice problems for each target problem. The three training problems followed the same pattern as the target problem; see Table 2. The training problems were constructed in such a way that they did not overlap (i.e., did not share surface features) with each other or the target problem. The low variability practice group was trained on the first of the three training problems and the high variability practice group was trained on all three.

In addition, there were two sequence extrapolation tutorial booklets, one for short instruction (12 pages) and one for long instruction (14 pages). Both tutorials consisted of general instructions on how to find pattern sequences as well as detailed descriptions of the component relations of patterns (e.g., forward, backward, repeat, and identity relations).

The long instruction tutorial had two additional pages of general instruction describing how to extrapolate patterns. Two example problems were worked through in detail, extrapolating one letter at a time. Participants were told that in order to extrapolate a pattern, they must first find the relations that make up the pattern, and then use those relations to continue it. For example, to extrapolate the first letter of the pattern below participants were instructed as follows:

A B M C D M . . . . → E F M

> "So first we need to decide on the 1st letter of the third period. The 1st letter in the second period, C, is one forward from the 2nd letter of the first period. So the letter we are looking for should be the one forward from the 2nd letter of the second period, which is D. So the letter we are looking for is E."

The rest of the extrapolation was described in a similar way. Participants were then given step-by-step extrapolation instructions for a more difficult problem. These instructions were intended to focus the learner on the procedure of extrapolating a pattern.

Both instruction groups also received a general tutorial test that consisted of four recall questions and one comprehension question for the short instruction group and two comprehension questions for the long instruction group. An example of a recall question was to write a brief description of the repeat letter relation. An example of a comprehension question was to describe what a period is and give an example of a periodic sequence. In addition, there were two diagrammatic illustrations of the underlying pattern relations for each of the target problems as well as two blank diagrammatic recall sheets (see Figure 5 for an example of a diagrammatic pattern illustration). There were also two distractor tasks that consisted of three multiplication problems each.

Test problems and training problems were presented via the same PsyScope computer system that was used in experiment 1. Direct instruction training materials were presented in booklet form.

Problem 1 and problem 2 and associated training stimuli were counter-balanced across all conditions.

Design

The participants were randomly assigned to one of four groups: *low variability practice* (*n* = 30), *high variability practice* (*n* = 31), *short instruction* (*n* = 28), and *long instruction* (*n* = 30). In addition, the *no-training group* (*n* = 30) from experiment 1 was included in the data analysis as a comparison group.

In both of the practice groups, participants solved letter sequence extrapolation problems that conformed to the same patterns as those used in the target problems. The low variability practice group solved one and the same training problem three times and the high variability practice group solved three training problems, each having different surface features (letters).

In the instruction training conditions participants first read general tutorials, then memorized and recalled the abstract patterns for each target problem. The only difference between the short and long instruction training was that the long instruction participants were given two additional pages in the tutorial that provided specific step by step instructions for how to extrapolate a problem; see Materials section.

Procedure

Participants were tested in groups of 1-4 people. The procedure consisted of two cycles, each encompassing a training phase and a test phase.

Procedure for practice groups. Participants were first given general instructions on how to solve sequence extrapolation problems. They were then given the first sequence extrapolation training problem. Participants were instructed to extrapolate each of the eight positions by clicking the mouse on any given position and typing in the answer; see Figure 2 for an example. They were given 6 minutes to solve each problem. After participants finished solving a problem or 6 minutes time elapsed, they continued to the next problem by pressing the space bar.

After participants solved all three training problems they were given the target problem instructions. Target problem instructions were the same as the training instructions except that they included the hint that if participants noticed a pattern on any of the prior problems it would help them solve the next problem. Participants were then given 6 minutes to solve the target problem. Finally, they were given the transfer problem and were instructed to solve it in the same manner as the target problem.

The second cycle proceeded in the same way. The entire procedure took 60-80 minutes.

Procedure for direct instruction groups. Before the training phase all participants were given the general tutorial text to read (maximum time allowed 18 minutes) after which they were given the tutorial test (maximum time 6 minutes). At the beginning of the training cycle participants were given 3 minutes to memorize the first diagrammatic pattern illustration. Next, participants were presented with the diagrammatic blank recall sheet and instructed to recall and write down the relations of the pattern (maximum time 3 minutes). Participants were then given the distractor task to prevent memory rehearsal strategies of the rules (maximum time 3 minutes). Next, participants were presented with the general instructions for the test problems. They were then given 6 minutes to solve the target problem. Finally, they were given the transfer problem and were instructed to solve it in the same manner as the target problem. The second cycle proceeded in the same way. The entire procedure took 70-90 minutes.

Results

The two primary questions are whether participants acquired knowledge of the target pattern from the training procedures, and whether that knowledge facilitated performance on the subsequent test problems (target and transfer). The question of the nature of the differences in what was learned is discussed in more detail after the report of experiment 3.

Alpha was set to .05 for all statistical tests and effect sizes (eta squared, η^2) were calculated for main effects, interactions and main comparisons. Cohen (1988; see also Olejnik & Algina, 2000) has suggested that effects be regarded as small when $\eta^2 < .06$, as medium when $.06 < \eta^2 < .14$ and as large when $\eta^2 > .14$.

Training Performance

Training Time

The average time spent on training was calculated for each training group. The low variability practice group spent 1,117 seconds (~19 minutes; SD = 412) on solving the practice problems whereas the high variability practice group spent 1,163 seconds (SD = 352). In contrast, the short instruction group spent 1,572 seconds (~26 minutes; SD = 197) and the long instruction group 1,624 seconds (SD = 150) on training (i.e., tutorial, test, memorization and recall of the abstract pattern). These data show that the instruction groups spent considerably more time than the practice groups with the training materials. If time on task is a major determinant of learning outcomes, we should therefore expect the instruction groups to perform better than the practice groups.

Practice

The first question is whether problem solving performance increased across practice trials. If participants extracted knowledge of the pattern from the initial problem it should facilitate performance on subsequent training problems. The low variability group should show a positive linear increase across training trials, because they solved the exact same problem on each occasion. The high variability group should show the same pattern of results if the knowledge gained from the initial problem was abstract and accessible for transfer across training problems.

The *practice training score* was determined in the same manner as the problem solving scores from experiment 1 as the number of correct extrapolations varying from 0-8. An initial one-way ANOVA was conducted to investigate the effect of problem (problem 1 vs. problem 2) on problem solving performance. The ANOVA revealed no effect of problem, *F* (1, 60) = 2.83, *MSE* = 3.41, *ns*, indicating that the type of pattern had no significant effect on problem solving performance.

Figure 6 shows the mean problem solving scores for the low and high variability groups on the three training problems collapsed across problem. A 2 (group: low variability vs. high variability) by 3 (trial: 1 vs. 2 vs. 3) mixed ANOVA was conducted to examine the effect of type of training across problem solving trials. The ANOVA revealed a large main effect of trial, *F* (2, 118) = 31.83, *MSE* = 1.45, $p < .05$, $\eta^2 = .35$, indicating that performance increased significantly with training. There was no effect of group, *F* (1, 59) = .87, *MSE* = 20.85, *ns*, indicating that participants in the low variability condition performed at the same level as participants in the high variability condition. There was also no interaction of group by trial, *F* (1, 118) = .87, *MSE* = 1.45, *ns*, indicating that performance did not significantly differ across training trials as a function of group.

To follow up the effect of trial a linear ANOVA was conducted on training trials 1-3. The ANOVA revealed a significant positive linear trend across training trials, *F* (1, 59) = 42.69, *MSE* = 1.82, $p < .05$, $\eta^2 = .42$, indicating performance improvements across all three trials collapsed across training groups.

Direct Instruction

Two training measures were taken: participant ratings of how well they understood the general tutorial and pattern recall. Participants were asked to rate how well they understood each page of the general tutorial. Ratings were made on a 1-5 Likert scale; from 1 (*I don't understand at all*) to 5 (*I understand completely*). Mean self-rating scores were 4.68 (*SD* = .48) for the short instruction group and 4.79 (*SD* = .43) for the long instruction group.

A one-way ANOVA comparing short instruction ratings to long instruction ratings was not significant, *F* (1, 57) = .75, *MSE* = .20, *ns*, indicating instruction groups were equally likely to report that they understood the general tutorial.

The second measure was the number of relations correctly recalled from the diagrammatic pattern descriptions. The *pattern recall score* was based on the number of relations embedded in a given pattern; hence, scores varied from 0-7 for pattern 1 and 0-8 for pattern 2. Figure 7 shows the proportion of correctly recalled pattern relations for short and long instruction groups for both patterns.

A 2 (group: short instruction vs. long instruction) by 2 (pattern: pattern 1 vs. pattern 2) mixed ANOVA revealed a large main effect for pattern, *F* (1, 56) = 39.05, *MSE* = .004, $p < .05$, $\eta^2 = .41$, indicating that participants recalled significantly more relations for pattern 1 than for pattern 2. There was no effect of group, *F* (1, 56) = .03, *MSE* = .11, *ns*, indicating that the two direct instruction groups did not differ in recall performance. However, there was a marginally significant interaction of group by pattern, *F* (1, 56) = 3.54, *MSE* = .004, *p* = .064, $\eta^2 = .06$, indicating that group recall performance differed as a function of pattern. Inspection of the means show that the short instruction group recalled slightly more relations than the long instruction group on pattern 1 (6.25 vs. 5.67) and the opposite trend was observed on pattern 2 (4.64 vs. 5.13).

We have no explanation for this unexpected finding. The effect was small, $\eta^2 = .03$ for pattern 1 and $\eta^2 = .009$ for pattern 2.

In sum, the two direct instruction groups scored high on the comprehension and recall tasks, indicating that they learned something about the relevant pattern. In addition, the two groups performed almost identically on both measures.

Test Performance

Accuracy

Participants were instructed that their task was to extrapolate the letters correctly. They were not instructed to work as fast as possible. Therefore, accuracy is the appropriate measure of task performance. The *problem solving score* was determined in the same manner as in experiment 1 (i.e., scores varied between 0-8).

An one-way repeated measures ANOVA was conducted to investigate effect of problem (1 vs. 2) on problem solving performance. The ANOVA revealed no effect of problem, *F* (1, 148) = .08, *MSE* = 2.14, *ns*, indicating that the type of pattern had no significant effects on problem solving performance. Columns 2 and 4 in Table 3 show the mean problem solving scores and standard deviations for the practice, direct instruction, and no-training groups on target and transfer problems collapsed across problem.

A 5 (training: low variability practice vs. high variability practice vs. short instruction vs. long instruction vs. no-training) by 2 (test problem: target vs. transfer) mixed ANOVA revealed a medium-sized main effect of training, *F* (4, 144) = 4.75, *MSE* = 14.49, $p < .05$, $\eta^2 = .12$, indicating that problem solving performance significantly differed across training groups. There was also a large main effect of test problem, *F* (1, 144) = 23.65, *MSE* = 1.50, $p < .05$, $\eta^2 = .14$, showing that participants performed significantly better on the target problem than on the transfer problem. There was no interaction of training by test problem, *F* (4, 144) = .67, *MSE* = 14.49, *ns*, indicating that the difference in problem solving performance between target and transfer problems did not vary as a function of training.

Follow up comparisons for training revealed that the low variability practice group performed significantly better than the short instruction group on problem solving performance, *F* (1, 144) = 9.32, *MSE* = 14.49, $p < .05$, $\eta^2 = .06$. Planned comparisons also revealed that the low variability practice group did not differ significantly from either the high variability practice group or the long instruction group, *F* (1, 144) = .52, *MSE* = 14.49, *ns*, and *F* (1, 144) = 1.21, *MSE* = 14.49, *ns*. The high variability practice group also did not significantly differ from the long instruction group, *F* (1, 144) = .09, *MSE* = 14.49, *ns*. The long instruction group also performed significantly better than the short instruction group, *F* (1, 144) = 4.20, *MSE* = 14.49, $p < .05$, $\eta^2 = .03$. In addition, the short instruction group did not significantly differ from the no-training group, *F* (1, 144) = .17, *MSE* = 14.49, *ns*. All comparisons that were significant for the entire participant group were also significant in the upper 2/3 analysis but with stronger effects.

We repeated the above analyses on the scores of those participants who received the top two-thirds of the scores on the training measures (see Table 3, columns 3 and 5). We refer to this as *the upper 2/3 analysis*. The reason for the upper 2/3 analysis is that the participants differed in their ability to learn from the various training procedures, presumably due to differences in cognitive ability, differences in motivation, and perhaps other factors. For the purpose of the present study *we are concerned with the effects of training on subsequent problem solving when training is successful*. Hence, participants who failed to learn during training, regardless of reason, dilute the effects we are interested in. The effect of restricting our attention to the upper 2/3 of the participants (based on the training measures, not test performance) ought to accentuate those effects that are due to differences between the types of training. The upper 2/3 analysis should therefore show the same pattern of effects as the analysis of the entire sample but with larger effects. As columns 3 and 5 in Table 3 show, this was indeed the case. For example the main effect of training increased from $\eta^2 = .12$ to $\eta^2 = .32$, similarly the effect of the low variability group over the short instruction group increased from $\eta^2 = .06$ to $\eta^2 = .12$ and the long instruction group over the short instruction group increased from $\eta^2 = .03$ to $\eta^2 = .07$. Any statistical test that was significant in the entire sample was also significant in the upper 2/3 analysis.

In addition, we were interested in whether or not problem solving performance differed across experiment cycles (cycle 1 vs. cycle 2). It is possible that participants became aware of the connection between the training and test phase after finishing cycle 1 and deliberately changed their approach to the training materials on cycle 2 thus leading to performance differences across cycles as a function of training type. A 4 (group: low variability practice vs. high variability practice vs. short instruction vs. long instruction) by 2 (cycle: 1 vs. 2) mixed ANOVA revealed an overall improvement across cycles for all training groups (F (1, 115) = 7.84, MSE = .2.01, $p < .05$, $\eta^2 = .06$). However, there was no interaction of training by cycle, indicating that improvement across cycles did not change as a function of training (F (3, 115) = .78, MSE = .2.01, *ns*). Performance improved uniformly across cycles for all training groups, suggesting that the participants benefited in some general way from their familiarity with the task domain but that they did not change their training strategies for cycle 2 in such a way as to interact with training type.

In sum, the accuracy results show that participants in the two analogy groups and the long instruction group performed better than the participants in the short instruction and no-training groups on both target and transfer problems. The long direct instruction and the practice groups did not differ from each other, thus constituting alternative paths to the same level of mastery. In addition, all groups performed better on the target than on the transfer problems.

Time Measures

We defined *solution time* as the total amount of time (in seconds) it took the participants to solve the problem. Following a long-standing tradition in cognitive psychology, we interpret a solution time as an estimate of the amount of cognitive processing the participant engaged in to produce his or her problem solution. A detailed analysis of solution times is presented after experiment 3. The purpose of the analysis reported here is to determine whether those participants who had higher accuracy scores engaged in more processing.

Table 4 shows the mean solution times and standard deviations for the practice, direct instruction, and no-training groups on target and transfer problems; see columns 2 and 4. As with accuracy, we also report the corresponding measures for those participants who received the upper 2/3 of the scores on the training measures; see columns 3 and 5. The same logic applies: Effects that are due to the training procedures ought to be accentuated when we restrict the analysis to the upper 2/3 of the participants.

A 5 (training: low variability practice vs. high variability practice vs. short instruction vs. long instruction vs. no-training) by 2 (test problem: target vs. transfer) mixed ANOVA revealed a large main effect of training, *F* (4, 144) = 20.15, *MSE* = 2.40, $p < .05$, $\eta^2 = .36$, indicating that solution time differed significantly across training groups. There was also a medium-sized effect of test problem, *F* (1, 144) = 15.01, *MSE* = .66, $p < .05$, $\eta^2 = .09$, showing that participants solved the target problem significantly faster than the transfer problem. In addition, there was a medium-sized interaction of training by test problem, *F* (4, 144) = 3.43, *MSE* = .66, $p < .05$, $\eta^2 = .09$, indicating that the solution time differed across test problems as a function of group.

Follow up comparisons revealed that the low variability practice group solved the target and transfer problems faster than both the short and long instruction groups, *F* (1, 144) = 14.37, *MSE* = 2.40, $p < .05$, $\eta^2 = .09$ and *F* (1, 144) = 29.93, *MSE* = 2.40, $p < .05$, $\eta^2 = .14$ respectively. Planned comparisons also revealed that the low variability practice group did not significantly differ from the high variability practice group, *F* (1, 144) = 3.30, *MSE* = 2.40, *ns*. In addition, the short and long instruction groups did not significantly differ from the no-training group, *F* (1, 144) = 1.89, *MSE* = 2.40, *ns* and *F* (1, 144) = .13, *MSE* = 2.40, *ns*.

Follow up comparisons for the interaction of training by test problem revealed that both practice groups and the long instruction group solved the target problems faster than the transfer problems, *F* (1, 144) = 5.93, *MSE* = .66, $p < .05$, $\eta^2 = .04$, *F* (1, 144) = 13.36, *MSE* = .66, $p < .05$, $\eta^2 = .08$, and *F* (1, 144) = 9.54, *MSE* = .66, $p < .05$, $\eta^2 = .06$. In contrast, the short instruction and no-training groups solved both target and transfer problems equally fast, *F* (1, 144) = .60, *MSE* = .66, *ns* and *F* (1, 144) = .15, *MSE* = .66, *ns*.

All statistical tests that are significant for the entire sample are also significant when the analysis is restricted to the upper 2/3 of the participants. As expected, the effects are accentuated; see Table 4, columns 3 and 5. For example, the main effect of training increased from $\eta^2 = .36$ to $\eta^2 = .43$, similarly the effect of the low variability practice group over the short and long instruction groups increased from $\eta^2 = .09$ to $\eta^2 = .11$ and $\eta^2 = .14$ to $\eta^2 = .17$ respectively.

In sum, the results show that the practice groups solved the target and transfer problems faster than both of the direct instruction and no-training groups. Although the long instruction group performed at the same level as the practice groups as measured by accuracy, they did not show an advantage over the short instruction and no-training groups with respect to solution time. In addition, both the practice and long instruction groups showed a significant slow down on the transfer problem.

Discussion

Direct instruction and practice can both facilitate performance on sequence extrapolation problems, as shown by the fact that performance on the target problem was more accurate for both the long instruction and the two practice groups than for the no-training group. This result was not self-evident, because not every training procedure produces higher accuracy in this domain: The short direct instruction group did not perform significantly better than the no-training group. The higher accuracy of the better groups was not due to time on task, because time spent on training was more for instruction than for practice, and the mean solution time on the target task was shorter for the groups with the higher accuracy. We infer that a person can master letter sequence extrapolation problems via either instruction or practice, but that what is learned differs in the two scenarios.

Type of knowledge. The fact that the two instruction groups took longer to complete the problems than the two practice groups is consistent with the expectation that the instruction scenario prompts the construction of more declarative than procedural knowledge, while the opposite is true for the practice scenario. Declarative knowledge needs to be proceduralized or compiled in order to guide the solving of an unfamiliar problem, a cognitive process that is complex and likely to take time. Practice, on the other hand, generates procedural knowledge and its application to a new problem is fast; hence the shorter solution times. (After the presentation of experiment 3, we show that this difference in time holds at each level of accuracy, but only for one component of the solution process.)

There was a clear difference between the short and long instruction groups in favor of the latter. The only difference in training between those two groups was the inclusion of two extra pages describing the type of inferences one needs to carry out to extrapolate a letter sequence. Because this information was procedural in nature, one might expect it to facilitate the generation of procedural knowledge. However, the lack of effect on the solution time is inconsistent with this prediction. Nevertheless, the difference in accuracy and solution time between the two groups demonstrate that the long instruction group benefited from the extra instruction in some way. It is possible that the two extra pages resulted in a *declarative* representation of the required inference type, but it is not clear why this representation would require as much cognitive processing to proceduralize (as shown by the long solution times) and nevertheless be less error prone (as shown by the high accuracy scores). We return to this issue in a later section.

Abstraction. The fact that the training facilitates performance on the target problem shows that what is learned during training is not completely specific. Although the target problem used the same pattern as the participants encountered during training, that pattern was instantiated in different letters. What was learned must therefore have been of at least intermediate abstraction in order to apply to the target problem. The fact that all groups took longer to solve the transfer problem shows that the abstraction level nevertheless was limited. Additional cognitive processing was needed to apply what had been learned to the transfer problem. (The longer solution times on the transfer problem are unlikely to be due to any difference in difficulty between the problems themselves, because the performance of the no training group revealed no such differences in difficulty; see experiment 1.)

The fact that there was no difference between the high and low variability groups with respect to solution time on either test problem is surprising in the light of prior research that claimed that variability in practice problems prompts abstraction. Recall that the low variability group solved exactly the same practice problem three times before encountering the target problem, while the high variability group solved three different problems. One would expect faster solution times for the high variability group on the target problem and less of a performance decrement on the transfer problem, but the high group shows no such advantages with respect to the low group. Because this lack of difference between the low and high variability groups is at variance with prior research, we replicate it in experiment 3.

Our confidence in the above conclusions is strengthened by the upper 2/3 analysis. Each of the effects mentioned above recurs in that analysis, but with a larger effect size. For example the main effect for training increased from $\eta^2 = .12$ to $\eta^2 = .32$ for accuracy and $\eta^2 = .36$ to $\eta^2 = .43$ for solution time. This is what one would expect if the effects are due to differences between the training conditions. After reporting experiment 3, we outline a processing account of our key findings.

EXPERIMENT 3: PRACTICE VS. INCIDENTAL TRAINING

Although direct instruction and practice are typical of formal learning situations, one often hears the statement that people learn best from 'experience.' What is meant by this is approximately that people extract information from situations they participate in, even in the absence of any intention to learn. In educational contexts, the belief in such experiential, unintentional learning implies instructional tactics that put students in situations that in some sense 'embodies' the target subject matter (manipulatives, microworlds). The hope is that they will extract the relevant knowledge from their interactions with those situations. In laboratory contexts, this idea has been researched under the two labels incidental learning and implicit learning. The main purpose of the laboratory studies have been to document the existence of this type of learning and they have succeeded in doing so (Berry, 1997; Reber, 1989; Stadler, 1998; Watternmaker, 1999).

However, the literature lacks detailed comparisons between what is learned under incidental training conditions and what is learned from more deliberate forms of training. The first purpose of experiment 3 was to compare incidental training with the practice scenario from experiment 2. In the incidental condition, the participants memorized letter sequences that instantiated one and the same pattern, without being told of this fact and without being told that the sequences were relevant for a subsequent problem solving task. Following the same strategy as in experiment 2, we implemented two parametric variants of the incidental training scenario, called *short* and *long*. They differed only with respect to the number of instances memorized (6 vs. 18).

The second purpose of experiment 3 was to replicate the unexpected lack of differences between the high and low variability practice conditions in experiment 2. Participants in the two practice groups received the exact same training as in experiment 2.

Method

Participants

Ninety-four undergraduate students from the University of Illinois at Chicago participated in return for course credit.

Materials

The test problems and the practice problems were the same as those used in experiment 2; see Tables 1 and 2. In addition, there were a total of 36 training strings consisting of 12 letters for problem 1 and 14 letters for problem 2, 18 strings for each target problem. The 18 strings associated with each target problem followed the same pattern as the given sequence for that problem; see the Appendix for examples. The short incidental group was trained on 6 strings per problem. The long incidental group was trained on 18 strings per problem. All participants received the two target and the two transfer problems. Target and transfer problems and their associated training stimuli were counter-balanced across all conditions.

Design

The participants were randomly assigned to one of four training groups: *low variability practice* ($n = 26$), *high variability practice* ($n = 23$), *short incidental* ($n = 25$), and *long incidental* ($n = 23$). In addition, the *no-training group* ($n = 30$) from experiment 1 was included in the statistical analyses as a comparison group.

In both of the practice groups, participants solved letter sequence extrapolation problems that conformed to the same patterns as those used in the target problems. As in experiment 2, the low variability practice group solved one and the same training problem three times and the high variability practice group solved three training problems, each having different surface features (letters). In both of the incidental learning groups, participants memorized letter strings that conformed to the same patterns as those used in the target extrapolation problems. The short incidental group memorized and recalled 6 training strings and the long incidental group memorized and recalled 18 training strings.

Procedure

Participants were tested in groups of 1-4 people. Each participant was tested on a Macintosh computer with a 17" color monitor, standard keyboard and mouse. All stimuli were presented in black 30 pt font in the center of the screen. The experiment was designed and presented with the PsyScope software.

The procedure consisted of two cycles. Each cycle was composed of training followed by solving one target and one transfer problem.

Procedure for practice groups. The procedure was exactly the same as in experiment 2.

Procedure for incidental learning groups. Participants were first instructed to memorize and recall each letter string one by one (6 strings for the short incidental group and 18 strings for the long incidental group). They were given 45 seconds to memorize each string and 30 seconds to recall and type in the string. After they finished recalling the string or 30 seconds time elapsed participants were presented the next string. This procedure was continued through all of the training strings. Participants were then instructed to describe the pattern in the memorization strings as best they could.

Next, participants were given instructions on how to solve the sequence extrapolation problems. Instructions included the hint that if the participants noticed a pattern from the training strings it would help them solve the sequence extrapolation problem. They were then presented with the target problem. They were given 6 minutes to solve the problem. Finally, they were given the transfer problem and were instructed to solve it in the same manner as the target problem.

To investigate how aware the participants were of the knowledge they acquired during training, they were asked at the end of each training cycle whether or not they noticed a pattern in the training strings. If they noticed a pattern they were asked to write down a description of it.

The second cycle proceeded in the same way. The entire procedure took 70-90 minutes.

Results

Two key questions are whether participants acquired knowledge of the target pattern from the training procedures and whether that knowledge facilitated performance on the subsequent test problems.

Training Performance

Training Time

The average time spent on training was calculated for each training group. The low variability practice group spent 1,202 seconds (~20 minutes; SD = 382) on solving the practice problems whereas the high variability practice group spent 1,366 seconds (SD = 419). In contrast, the short incidental group spent ~ 900 seconds (~15 minutes) and the incidental long group spent ~ 2,160 seconds (~36 minutes). These data show that the low and high variability practice groups spent approximately the same amount of time on training whereas the short incidental group spent a shorter amount of time and the incidental long spent considerably more time with the training materials. If time on task is a major determinant of the learning outcomes, we should therefore expect the long incidental group to exhibit superior performance.

Practice

Did training performance increase across practice trials? Because we used the same training procedure in this experiment as in experiment 2, we expected to find positive linear trends across trials but no effect of group. The *practice training score* was determined in the same manner as the problem solving scores from experiments 1 and 2. Because participants were asked to extrapolate each problem to eight positions their scores varied from 0 to 8.

An initial repeated measures one-way ANOVA was conducted to investigate the effect of problem (problem 1 vs. problem 2) on problem solving performance. The ANOVA revealed no effect of problem, F (1, 48) = 1.98, MSE = 4.34, *ns*, indicating that the type of pattern had no effect on analogy training performance. Figure 8 shows the mean problem solving scores for the low variability and high variability practice groups on the three training trials collapsed across problem.

A 2 (group: low variability vs. high variability) by 3 (trial: 1 vs. 2 vs. 3) mixed ANOVA was conducted to examine the effect of training group across problem solving trials. The ANOVA revealed a large main effect of trial, F (2, 94) = 15.35, MSE = 1.76, $p < .05$, $\eta^2 = .25$, indicating that performance significantly differed across problems. There was no effect of group, F (1, 47) = .02, MSE = 18.39, *ns*, indicating that participants in the low variability group performed at the same level as the participants in the high variability group in experiment 2. There was also no interaction of group by trial, F (2, 94) = .442, MSE = 1.76, *ns*, indicating that performance did not significantly differ across training trials as function of group.

To follow up the effect of trial a linear ANOVA was conducted on training trials 1-3. The ANOVA revealed a significant positive linear trend across training trials, F (1, 47) = 24.70, MSE = 2.16, $p < .05$, $\eta^2 = .35$, indicating performance improvements across all three trials collapsed across training groups. Similar to experiment 2, these results show that both the low and high variability training groups constructed knowledge of the pattern thus facilitating performance on the second and third problem trials.

Incidental Training

The first question of interest is whether the participants acquired any knowledge of the pattern by memorizing the training sequences. Knowledge of the pattern can be used to reconstruct the letter sequence and should thus improve recall performance. If pattern knowledge is gradually acquired via memorization participants should perform better on later trials than on earlier trials[1].

The *incidental training score* was determined by the number of letters correctly recalled in the memorization task. Because there were 12 letters to memorize for pattern 1 and 14 letters to memorize for pattern 2, the memory scores varied from 0 to 12 and 0 to 14 respectively. Mean memory scores for both short and long training groups on patterns 1 and 2 are presented in Figure 9.

To investigate whether or not participants recall scores' increased across trials linear ANOVAs were conducted on trials 1-6 for the short incidental group and trials 1-18 for the long incidental group for both patterns 1 and 2. The short incidental group showed a positive linear trend for pattern 1 but not for pattern 2, $F\ (1, 24) = 5.94$, $MSE = 10.85$, $p < .05$, $\eta^2 = .20$ and $F\ (1, 24) = .16$, $MSE = 10.23$, *ns* respectively. This result indicates that participants in the short incidental training group acquired some knowledge of the pattern sequence for pattern 1 but not for pattern 2. In contrast, the long incidental group showed large positive linear trends for both patterns 1 and 2, $F\ (1, 22) = 39.42$, $MSE = 10.85$, $p < .05$, $\eta^2 = .64$ and $F\ (1, 22) = 22.30$, $MSE = 10.61$, $p < .05$, $\eta^2 = .50$. These results indicate that participants in the long incidental training group acquired knowledge of both patterns from the memorization training.

In addition, a 2 (group: short incidental vs. long incidental) by 2 (pattern: pattern 1 vs. pattern 2) mixed ANOVA was conducted to investigate whether pattern knowledge increased as a function of amount of training. The ANOVA revealed a large main effect of training group, $F\ (1, 46) = 7.91$, $MSE = .003$, $p < .05$, $\eta^2 = .15$, indicating that the long incidental group performed significantly better than the short incidental group. There was no effect of pattern, $F\ (1, 46) = 1.65$, $MSE = .002$, *ns*, and no interaction of training group by pattern, $F\ (1, 46) = .41$, $MSE = .002$, *ns*. These results show that the long incidental group recalled significantly more letters than the short incidental group on both patterns 1 and 2.

To investigate how aware the participants were of the knowledge they acquired during training, the participants' written pattern descriptions were coded by two independent coders with a reliability of 85%. All disagreements were arbitrated by the first author. The pattern descriptions were coded using both a strict and loose coding scheme.

The *strict coding scheme* was based on the number of specific relations the participants identified in the pattern; see Figure 1. Hence, the strict coding score varied from 0-7 relations for pattern 1 and 0-8 relations for pattern 2. The *loose coding scheme* was based upon demonstration of general pattern concepts such as forward relations, backward relations, repeat, and period.

[1] In a previous study (Nokes & Ohlsson, 2000) we showed that participants who memorized and recalled *random* symbol sequences did not show benefits in performance across trials.

Figure 10a shows the proportion of pattern relations identified (i.e., strict coding) by the short and long incidental groups for both patterns. A 2 (group: short incidental vs. long incidental) by 2 (pattern: pattern 1 vs. pattern 2) mixed ANOVA revealed no effects for group or pattern, *F* (1, 46) = 1.49, *MSE* = .044, *ns* and *F* (1, 46) = 1.45, *MSE* = .06, *ns* respectively. In addition, there was no interaction of group by pattern, *F* (1, 46) = .13, *MSE* = .044, *ns*. These results indicate that short and long incidental groups did not differ in their identification of the relations for either pattern. Overall, pattern identification was quite poor showing that both groups identified not more than approximately 12% of the specific pattern relations.

Furthermore, an analysis of the pattern descriptions for general concepts revealed no effect of training group, *F* (1, 46) = 1.45, *MSE* = .06, p = .074, indicating that the long and short incidental groups performed similarly on pattern concept descriptions. Figure 10b shows the proportion of general concepts correctly identified for both groups on each pattern.

These results suggest that overall incidental training participants had very limited explicit knowledge of the patterns embodied in the memorization strings.

Test Performance

Did the knowledge gained from training facilitate performance on the target and transfer problems? Was the behavior of the incidental group similar to that of the practice groups, the direct instruction groups, or different from both?

Accuracy

The *problem solving score* was determined in the same manner as experiment 1 and 2 (i.e., scores varied between 0-8). An initial one-way repeated measures ANOVA was conducted to investigate effect of problem (1 vs. 2) on problem solving performance. The ANOVA revealed no effect of problem, *F* (1, 126) = 2.41, *MSE* = 2.95, *ns*, indicating that the type of pattern had no significant effects on problem solving performance. Table 6 shows the mean problem solving scores for the practice, incidental, and no-training groups on target and transfer problems collapsed across problem.

A 5 (training: low variability practice vs. high variability practice vs. short incidental vs. long incidental vs. no-training) by 2 (test problem: target vs. transfer) mixed ANOVA revealed a medium-sized main effect of training, *F* (4, 122) = 4.10, *MSE* = 13.09, $p < .05$, $\eta^2 = .12$, indicating that problem solving performance significantly differed across training groups. There was also a medium-sized main effect of test problem, *F* (1, 122) = 16.06, *MSE* = 1.88, $p < .05$, $\eta^2 = .12$, showing that participants performed significantly better on the target problem than on the transfer problem. In addition, there was no interaction of training by test problem, *F* (4, 122) = .49, *MSE* = 1.88, *ns*, indicating that problem solving performance did not differ across training groups as a function of test problem.

Follow up comparisons for training revealed that the low variability practice group performed significantly better than the long incidental group on problem solving performance, *F* (1, 122) = 4.08, *MSE* = 13.09, $p < .05$, $\eta^2 = .03$. Planned comparisons also revealed that the low variability practice group did not significantly differ from the high variability practice group, *F* (1, 122) = .03, *MSE* = 13.09, *ns*, and the long incidental group did not significantly differ from the short incidental or no-training groups, *F* (1, 122) = .77, *MSE* = 13.09, *ns* and *F* (1, 122) = .13, *MSE* = 13.09, *ns* respectively.

To address the question as to whether or not participants became aware of the connection between training and test we compared test performance on cycle 1 to test performance on cycle 2. A 4 (group: low variability practice vs. high variability practice vs. short incidental vs. long incidental) by 2 (cycle: 1 vs. 2) mixed ANOVA revealed an overall improvement across cycles for all training groups (F (1, 93) = 10.00, MSE = 3.01, $p < .05$, $\eta^2 = .09$). However, there was no interaction of training by cycle indicating that improvement across cycles did not change as a function of training type (F (3, 93) = .56, MSE = 3.01, *ns*). In sum, these results indicate that training strategies did not change for cycle 2.

In sum, the accuracy results show that the two practice groups performed significantly better than the two incidental and the no-training groups on target and transfer problems. Participants also performed better on the target than on transfer problems.

Solution Time

As in experiments 1 and 2, the *solution time* was the total amount of time (in seconds) it took the participant to solve the problem. Table 7 shows the mean solution times and standard deviations for the practice, incidental, and no-training groups on target and transfer problems.

A 5 (training: low variability practice vs. high practice vs. short incidental vs. long incidental vs. no-training) by 2 (test problem: target vs. transfer) mixed ANOVA revealed a large main effect of training, F (4, 122) = 9.90, MSE = 2.70, $p < .05$, $\eta^2 = .25$, indicating that solution times differed significantly across training groups. There was also a medium-sized main effect of test problem, F (1, 122) = 14.09, MSE = .69, $p < .05$, $\eta^2 =$ .10, showing that participants solved the target problem significantly faster than the transfer problem. In addition, there was no interaction of training by test problem, F (4, 122) = 1.40, MSE = .69, *ns*, indicating that solution times did not differ across training groups as a function of test problem.

Follow up comparisons for training revealed that the low variability practice group solved the test problems faster than the long incidental group, F (1, 122) = 8.10, MSE = 2.70, $p < .05$, $\eta^2 = .06$. Planned comparisons also revealed that the low variability practice group did not significantly differ from the high variability practice group, F (1, 122) = .92, MSE = 2.70, *ns*, and the long incidental group did not significantly differ from the short incidental group, F (1, 122) = 1.14, MSE = 2.70, *ns*. However, the long incidental group solved problems faster than no-training group, F (1, 122) = 5.87, MSE = 2.70, $p < .05$, $\eta^2 = .05$, whereas short incidental condition did not, F (1, 122) = 1.80, MSE = 2.70, *ns*.

In sum, the solution time results show that the practice groups solved the target and transfer problems faster than the incidental and no-training groups. The long incidental group was faster than the no-training group. The participants solved the target problems faster than the transfer problems.

Discussion

What did the incidental training groups learn? The participants in the incidental training scenario had the opportunity to extract knowledge of the relevant patterns, but no opportunity to practice extrapolation inferences. One might therefore expect them to exit the training with knowledge structures similar to those of the direct instruction groups from experiment 2: Declarative knowledge of the pattern, but no relevant procedural knowledge. This should generate higher accuracy scores than the no-training group, combined with long solution times.

This is not the pattern we observed. The accuracy scores provide no evidence that the incidental training groups could draw upon their knowledge of the pattern to improve their problem solving performance. In fact, the only evidence that the incidental training participants benefited from the training is the fact that the long incidental group had shorter mean solution time than the no-training group. In light of the fact that the solution time for the long incidental group increased rather than decreased when we restricted the analysis to the upper 2/3 of the participants (see Table 7), we are inclined to discount that single significance test and conclude that the test performance of the incidental training groups did not benefit from the training.

This conclusion does not imply that the incidental training participants did not learn anything. The fact that both incidental groups improved their recall scores across training trials, with stronger effect for the long group, is evidence that they gradually extracted the relevant pattern from the training sequences. The conclusion is not that they did not learn during training, but that they learned something that they could not use in subsequent problem solving. This conclusion is consistent with our previous findings (Nokes & Ohlsson, 2000).

A plausible interpretation of this outcome is that the incidental training participants acquired knowledge *without any associated retrieval structure.* That is, a relational structure that captured the regularities in the memorization sequences was gradually forming in memory, but the participants had no way of retrieving that schema for use in any other task than the memorization task for which it was created in the first place.

Due to the lack of effects on test performance, we cannot pursue the questions of what type of knowledge this was and at what level of abstraction it might have been encoded.

Replication of practice results. The two practice groups in this experiment behaved similarly to the corresponding groups in experiment 2. They exhibited both high accuracy and fast solution times. In particular, the low and high variability variants did not differ from each other on either target or transfer problems. The expected effect of variable practice on the level of abstraction once again failed to appear.

Our confidence in these conclusions are once again strengthened by the upper 2/3 analysis. All effects that are significant in the entire sample are also significant in the upper 2/3 analysis, but the effects are larger (with the exception noted above). For example, the main effect of training increased from $\eta^2 = .12$ to $\eta^2 = .29$ for accuracy and $\eta^2 = .25$ to $\eta^2 = .29$ for solution time. In addition, the effects of the low variability practice group over long incidental group increased from $\eta^2 = .03$ to $\eta^2 = .09$ for accuracy and $\eta^2 = .06$ to $\eta^2 = .08$ for solution time.

MEASURES OF TRANSFER

Up to this point, we have used performance on the transfer problem as our indicator of how well the participants could apply what they had learned to a novel problem. However, transfer is traditionally quantified in terms of *the relative advantage of a training group over a control group* (Singley & Anderson, 1989). That is, how much better does a training group perform than a no-training control group? The purpose of this section is to report relative advantage measures for all learning scenarios studied in Experiments 2 and 3.

In order to quantify the amount of transfer observed from the training tasks to the test tasks we calculated two measures of transfer, adapted from Singley and Anderson (1989), for both the target and the transfer problems; see Table 7 for a summary of transfer effects based on accuracy performance and Table 8 for a summary of transfer effects based on solution times.

The first transfer measure quantifies the amount of improvement of the experimental group (training group) over the control group (no-training group) normalized by control group performance, $T_{\% \text{ Improvement}} = (E - C / C)$ x 100; see concept 1 in Table 3 and 4. This transfer measure revealed large amounts of transfer for the practice and long instruction groups on accuracy performance revealing ~50% improvement on the target and ~60% improvement on the transfer problems. In contrast, the incidental and short instruction groups showed small to nil transfer effects for both target and transfer problems (less than 13% improvement).

We also calculated this measure for the upper two-thirds of training participants and found even larger effects for practice and long instruction groups showing ~98% and ~70% improvement respectively on the target task and ~98% and ~96% improvement on the transfer task. Similarly the upper two-thirds of incidental and short instruction groups also showed increased transfer effects revealing ~19% and ~20% improvements respectively on both target and transfer. However, these effects still remain considerably smaller than that of the practice and long instruction groups.

The second transfer measure quantifies the amount of transfer given the total amount of learning possible; $T_{\% \textit{ Total Learning}} = (E - C / {}_{\textit{Perform limit}} - C)$ x 100; see concept 2 in table 7 and 8. If one knows maximum performance possible on the target task (in this case 8 correct extrapolations for accuracy) you can calculate the total amount of transfer possible (i.e., the performance limit minus control performance). This measure revealed large amounts of transfer for the practice and the long instruction groups on accuracy performance showing ~49% improvement on target and ~36% improvement on transfer problems. In contrast, the incidental and short instruction groups showed small transfer effects for both on target and transfer problems (less than 11% improvement).

We also calculated this measure for the upper two-thirds of training participants and found even larger effects for practice and long instruction groups showing ~80% and ~57% improvement respectively on the target task and ~62% and ~60% improvement on the transfer task. Similarly the upper two-thirds of long incidental and short instruction groups also showed increased transfer effects revealing ~14% and ~15% improvements respectively on both target and transfer, however these effects still remain considerably smaller than that of the practice and long instruction groups.

In short, the relative advantage measure of transfer supports the conclusions reached through the previous analyses. In both experiment 2 and 3, the practice groups showed large amounts of transfer on both target and transfer problem while neither the instruction groups in experiment 2 nor the incidental training groups in experiment 3 did. This analysis also added to the evidence that the outcome was split for the two instruction groups. The long instruction group exhibited almost as much transfer as the practice groups, but the short instruction group was far behind. In the next section, we move towards an explanation of these findings by considering in detail how much processing was required in each group.

TOWARDS A PROCESSING ACCOUNT

The descriptive analyses presented in the previous sections reveal that there is non-trivial structure in the outcomes. There are indeed multiple paths to mastery in the letter sequence extrapolation domain. Data from the training phase support the conclusion that all three scenarios produced knowledge about the pattern in the target problem. The instruction groups could recall a high percentage of the relations in the patterns, particularly for pattern 1 (see Figure 7); the practice groups improved in accuracy across their three practice problems (see Figures 6 and 8); and the incidental training groups showed an increase in accuracy across their memorization trials (see Figure 9). (This trend is only strongly supported for the long incidental group.) In short, the participants in all three scenarios learned something about the target pattern.

The long instruction group and the two practice groups could utilize their knowledge about the pattern to perform better than the control group, as measured by mean accuracy on the target problem. Recall that these two scenarios were quite different. The long instruction group engaged in deliberate study of the target pattern. They acquired some lexical concepts to help them think about the relations in the pattern, saw a graphical display of the pattern, and attempted to recall the individual relations in the pattern. They did not engage in any activity that required extracting an unfamiliar pattern from a novel letter sequence, nor did they attempt to carry out any extrapolation inferences. The situation for the two practice groups was the opposite. After minimal instruction in the nature of the task, the participants in the practice groups solved three sequence extrapolation problems. They did not study any graphical or verbal representation of the pattern, nor did they engage in any discourse about patterns. These two learning scenarios differ qualitatively in terms of what information was available to the learners and how that information was presented. Yet, the participants in the long instruction group performed at approximately the same level as the participants in the two practice groups, as measured by mean accuracy on the target problem.

This outcome was not self-evident. Not all groups performed the target task at the same level of accuracy. Neither of the two incidental training groups performed better than the control group (see Table 5). With respect to direct instruction, the results were mixed. Although the long instruction group performed better than the control group, the short instruction group did not. The fact that different types of instruction can result in different levels of post-instruction performance is neither novel nor surprising. Such differences are traditionally explained by claiming that the learners in the different scenarios acquired different amounts of knowledge. One group performs less well than another because the members learned a subset of what another group learned.

An explanation in terms of amount of knowledge for those groups that performed differently would commit us to the belief that the groups that performed similarly acquired exactly the same knowledge. Given the qualitative differences between the instruction and practice scenarios, this is implausible. The purpose of this section is to formulate an account of what happened in our experiments in terms of what was learned – which type of knowledge and which level of abstraction – and in terms of the cognitive processing that the knowledge required in each learning scenario. Although we rely primarily on the accuracy measures in the previous sections, in this section we rely on the solution times, interpreted as measures of amount of cognitive processing.

Closer scrutiny of the solution times show that the three scenarios we studied differed in ways that cut across similarities and differences in mean accuracy. For this analysis, we divided the solution time into two additive parts, the *deliberation time*, the time until the participant typed in their first letter, and the *extrapolation time*, the time it took him or her to complete the problem. If we allow ourselves the reasonable assumption that the participants tried to figure out the pattern before they attempted to extrapolate it, then the deliberation time is an estimate of how long it took the participant to identify the pattern in the given sequence. The extrapolation time is an estimate of how long it took him or her to carry out the extrapolation inferences. We focus on those participants who performed in the upper two-thirds on the training task of their instructional condition. Table 8 shows the deliberation and extrapolation times for the control group and for each group in Experiments 2 and 3. In the following analysis, we focus on the long instruction group, the high variability practice group, and the long incidental training group, using the results from the parametric variants of these scenarios as auxiliary evidence. The purpose is to propose a story about what knowledge each group acquired that is consistent with the pattern of findings.

Interpretation: Direct instruction. During training, the long instruction participants learned to recall the relations in the pattern to a high degree of accuracy, so they must have acquired some representation of the target pattern. These participants had no opportunity to practice, so it is plausible that their knowledge consisted of a schema or some other type of declarative representation. This representation was applicable to a novel letter sequence, because the deliberation time for the target problem was shorter for the instruction group than for the control group (t (48) = 4.48, p < .05). Hence, it must have been of at least first-order abstraction. Presumably, the deliberation time for this group was spent matching their schema to the given sequence in the target problem.

The data suggest that their schema was also applicable to the transfer problem with little further cognitive work. The time to the first extrapolation was no longer for the transfer problem than for the target problem; instead, it was 9 seconds shorter (see Table 8; t (19) = .909, *ns*), a clear case of declarative transfer. Presumably, the facilitation occurred because the new pattern had the same structure, i.e., the important relations held between the same positions in the target and transfer patterns, and because the particular relations in the transfer pattern were generated from the relations in the target pattern via a systematic transformation (i.e., stretching; see Method section for Experiment 1). In sum, the data suggest that the deliberate study of the target pattern generated an abstract, easily transferable representation of the target pattern.

The extrapolation time for the direct instruction participants presents a contrasting picture. Because they had had no prior opportunity to practice extrapolation inferences during the training phase, the instruction participants – especially those who performed at a high level of accuracy – must have constructed the necessary procedural knowledge in the course of performing the target problem. Consistent with this, the time for extrapolation on the target problem was considerably longer in the long instruction group than in the practice groups; see below (t (39) = -7.23, $p < .05$ and t (38) = -5.15, $p < .05$, for high and low variability).

One might expect the procedural knowledge generated from the declarative schema to be at the same level of abstraction as the schema itself and hence easily transferable, but the data suggest that this was not the case. The extrapolation time for the target problem was 156 seconds, and for the transfer problem 214 seconds; see Table 8 (t (19) = -3.80, $p < .05$). This 37% increase contrasts with the slight decrease in the corresponding deliberation times. This observation is supported by the relative advantage measure of transfer, as applied to the solution times; see Table 9. As that table shows, the relative advantage of the long instruction group on the transfer task was close to zero, even slightly negative. The short instruction group had a slightly higher advantage, but not as high as the practice groups. This result should be considered in the context of short transfer distance from the target to the transfer problem. It appears that the amount of cognitive work required for extrapolation on the transfer problem (at which point the participants had some relevant procedural knowledge, constructed in the course of solving the target problem) was greater than the work needed for extrapolation on the target problem (for which the participants had no prior procedural knowledge). The participants either invested cognitive work into generalizing the rules created in the course of solving the target problem, or else they started over and created brand new rules for the transfer problem. Under either interpretation, we infer that the procedural knowledge constructed in the course of solving the target problem was not transferable.

This point deserves elaboration, because it is both counterintuitive and novel with respect to the literature on the interaction between declarative and procedural knowledge. The simulation models of knowledge compilation proposed by Anderson (1983) and Ohlsson (1996) implicitly assume that *knowledge compilation preserves abstraction level*. Briefly put, the resulting production rules have variables where the declarative structures they are derived from have variables. But in our experiments, the compilation of the abstract declarative schema into procedural knowledge appears to have produced less abstract knowledge. Although the deliberate understanding of the target schema helped the participants identify the pattern in the transfer task, the procedural knowledge of how to carry out the relevant extrapolation inferences in the target problem was not equally helpful in carrying out the corresponding inferences in the transfer problem. The knowledge compilation process began with an abstract, transferable schema but ended with a specific, non-transferable skill. We discuss this conclusion further below.

The short instruction group exhibited the same pattern of results as the long instruction group with respect to the deliberation and extrapolation times. The deliberation time on the target problem was shorter than that of the control group (t (47) = 4.02, $p < .05$) and comparable to that of the long instruction group. The extrapolation time was longer than for the high and low practice groups (t (38) = -7.14, $p <.05$ and t (37) = -5.19, $p <.05$), and again comparable to that of the long instruction group. Finally, the short instruction group showed the same pattern of a small decrease in deliberation time on the transfer problem, coupled with an extrapolation time that was as long as that for the target problem. Surprisingly, the short instruction group had slightly higher relative advantage than the long instruction group; see Table 9. We have a no explanation for this unexpected finding. Although the short instruction group did not perform as well as the long instruction group as measured by mean accuracy, the replication of the structure of deliberation and extrapolation times indicate that the processing was similar and that the knowledge this group acquired was of the same type and level of abstraction.

Interpretation: Practice. The high variability practice group improved its performance across the three practice problems. Because the practice participants were not shown the underlying pattern but had to work it out for themselves, this improvement most likely signals the acquisition of a mixture of declarative knowledge about the target pattern and procedural knowledge about how to detect patterns. This representation was more useful than the one learned by the instruction groups, because the deliberation time for the high variability practice group on the target problem was shorter than the corresponding times for either the long or the short instruction group; see Table 8, t (39) = -2.56, $p < .05$ and t (38) = -3.25, $p < .05$.

What was the level of abstraction attained by this group? The practice group exhibited an increase in time to first extrapolation from 32 seconds on the target problem to 48 seconds on the transfer problem, in contrast to the small decrease in deliberation time exhibited by both instruction groups. Hence, what the high-variability group learned about the pattern cannot have been so abstract that it applied to the transfer problem without modification. As one would expect, their knowledge was characterized by first-order abstraction.

Over the course of attempting three different extrapolation problems during the training phase of the experiment, the high variability practice group must have created the procedural knowledge needed to carry out the relevant extrapolation inferences. Because these participants saw three different letter sequences that all followed the same pattern, they had an incentive and opportunity to create abstract rules to carry out the extrapolation inferences. The data bear this out. The practice group exhibited a shorter extrapolation time on the target problem than the instruction group (59 versus 156 seconds; t (39) = -7.23, $p < .05$), indicating that they could apply the procedural knowledge they had already constructed rather than starting from scratch. Once again, this observation is supported by the relative advantage measure. The practice groups exhibited almost 50% transfer on this measure; see Table 9. The value was even higher for the upper 2/3 of the participants. They could achieve this in the span of only four prior problems (three practice, one target) because the distance between target and the transfer patterns was short.

If those rules were of intermediate abstraction, they needed to be revised and adjusted to apply to the transfer problem but the adjustment would be minor. Again, the data bear out this expectation. For the high variability practice group, the extrapolation time for the transfer problem was 43 seconds longer than the time for the target problem (t (20) = -8.10, $p < .05$).

There was reason to expect a different pattern to hold in the low variability practice group. Because they saw exactly the same problem (the same given letter sequence) three times before encountering the target problem, they had an opportunity to specialize their procedural knowledge to that particular sequence. If this had happened, we would expect to see evidence that they had to invest more cognitive work into adapting their pattern knowledge to the target problem (which for them exhibited a novel letter sequence) than the high variability group. This is what the data show. Time to the first extrapolation on the target problem was indeed longer for the low variability than for the high variability group, 56 seconds versus 32 seconds (t (39) = 2.23, $p < .05$). The same argument would predict that the time to compute the extrapolation inferences would also be longer. Once again, this is the case, 75 seconds versus 59 seconds for the high variability group. Once they abstracted their knowledge to fit the target problem, the low variability group should be at a smaller disadvantage on the transfer problem. This was indeed the case; the deliberation time on the transfer problem was still somewhat longer (66 seconds versus 48, but the extrapolation time was virtually the same as for the high variability group (96 seconds compared to 102).

This structure was replicated to a high level of detail in experiment 2. Once again, the low variability group exhibited slightly longer deliberation time than the high variability group on both the target and transfer problems, but the extrapolation times were virtually identical on both problems (60 versus 63 seconds for target, and 104 versus 102 seconds for transfer). In short, although the low variability group incurred an initial cognitive cost of the kind one would expect – they had to put more cognitive work into adapting their knowledge to the unfamiliar tasks than the high variability group – this initial disadvantage was overcome as soon as they had had an opportunity to abstract their procedural knowledge.

Comparing instruction and practice. The deliberation times for instruction and practice were comparable, but the extrapolation time was significantly shorter for the practice groups. This is entirely consistent with the idea that the latter had an opportunity to construct procedural knowledge while the former did not.

However, the magnitude of the extrapolation times for the transfer problems allows a further inference. If the instruction groups constructed procedural knowledge while solving the target problem, they then needed to transfer that knowledge to the transfer problem. But so did the practice groups. There is thus no obvious reason why the extrapolation times on the transfer pattern should differ between the two scenarios. In fact, the extrapolation time for the transfer problem was twice as long for the long instruction group as for the high variability practice group, 214 seconds as compared to 102 seconds (t (39) = -7.54, p < .05). The results from the short instruction group and the low variability practice group are in accord with this pattern; see Table 8. Further support comes from the relative advantage measure. The instruction groups showed no more relative advantage than the incidental learning group, but the practice groups were superior to both on this measure; see Table 9.

These results support the conclusion already reached previously that after carrying out the extrapolation inferences for the target problem, the instruction groups had to carry out significant cognitive work to transfer the resulting procedural knowledge to the transfer task. The implication is that whatever procedural knowledge they generated in the course of solving the target task was not abstract. This conclusion holds up even if we measure transfer as relative advantage vis-à-vis the control group; see Table 8. The two practice groups showed a consistently high relative advantage. The measure for the long instruction group is comparable, but for the short instruction group considerable lower.

Results separated by accuracy. The interpretation suggested above could be critiqued on the grounds that the deliberation and extrapolation times are likely to be imprecise estimates of the time needed to identify the pattern and to carry out the extrapolation inferences. In particular, participants might have oscillated between pattern detection and pattern extrapolation to a greater extent than our analysis presupposes. In particular, one could argue that the tendency to oscillate in this manner is a function of how well the person understands the relevant pattern, so that what looks like differences between the groups with respect to deliberation time and extrapolation time is in actuality nothing but a side effect of differences in performance level as measured by mean accuracy.

This argument is contradicted by Table 10. This table shows that the pattern of findings described above recurs at each level of accuracy. For this analysis, we divided the participants into three levels based on their accuracy on the target problem. The top level included participants who performed perfectly with 8 correct extrapolations, the second level those with 5-7 correct extrapolations, and the bottom level with 0-4 correct. To facilitate the comparison between accuracy levels, we limit Table 10 to the long instruction and high variability practice groups, the two groups for which our claim of alternative paths to mastery is strongest.

As Table 10 shows, those features of the data on which we based our interpretation recur at each level of accuracy. The deliberation times for the practice groups are comparable (with a slight advantage for practice in the top and bottom levels); the deliberation times for transfer are longer than target for practice but not for the instruction in the top and bottom levels; the extrapolation times for instruction are more than twice as long for practice on the target problem; and the instruction group does not catch up with the practice group on the target problem. At each of the three levels of accuracy, the extrapolation time for instruction is longer on the transfer than on the target problem. More importantly, at each level of accuracy, the extrapolation time is more than 100 seconds longer for instruction than for practice.

The recurrence of this pattern at each level of accuracy shows that the pattern is not a side effect of the overall differences in accuracy. In particular, those instruction participants who solved the target problem perfectly obviously did not have any serious deficiency in their knowledge and understanding of the relevant patterns nor any lack of relevant cognitive ability, but they still exhibited radically longer extrapolation times on both transfer and target than the practice groups. This indicates, first, that they had to invest cognitive work into compiling the required procedural knowledge, and, second, that the resulting procedural knowledge was not abstract. We discuss the implications of this further below.

Interpretation: Incidental training. The data from the long incidental training group shows that there was a significant increase in performance across the 18 memorization trials (see Figure 9). Thus, these participants learned something about the pattern. (In past studies, we have found that people do not improve when they are given random sequences to memorize; Nokes & Ohlsson, 2000.) Nevertheless, the incidental training scenario produced a pattern of results that does not conform to the pattern for either the instruction or the practice scenario.

Could the acquired representation be used to support performance on the subsequent target and transfer problems? The time to the first extrapolation for the long incidental group is more than two to three times as long on the target problem as the corresponding times for the long instruction and high variability practice groups, and comparable to the time for the no-training group. The time to the first extrapolation on the transfer problem is again longer than for the other groups. Recall also that the mean accuracy for the long incidental training group was not statistically different from the mean of the control group. In short, although studies of incidental and implicit learning have verified the existence of automatic pattern extraction during memorization (e.g., Reber, 1989), there is no evidence that the subjects in our incidental training scenario acquired a representation that they could utilize to support their problem solving performance.

Because the incidental training participants had no opportunity to practice sequence extrapolation, we predicted that they would came out of the training phase without any relevant procedural knowledge. Unexpectedly, their extrapolation time on the target problem is shorter than the time for the long instruction group, albeit longer than for the practice groups. On the transfer problem, the time for extrapolation is shorter for the long incidental group than for either the direct instruction group or either of the two high variability practice groups (from Experiments 1 and 2, respectively). We infer that the memorization condition produced procedural knowledge that was helpful with respect to the extrapolation steps. The effect on the extrapolation time for the transfer problem shows that this knowledge was transferable.

Given the difference between recalling a sequence and generating it by extrapolating a pattern, there is a question as to what the nature of this knowledge might be. A possible hypothesis is that the participants acquired a goal structure that helped them organize the recall of the memorized strings, and that this goal structure could be recruited to organize the extrapolation effort as well. However, it is not clear why this goal structure would be as useful as, or even more useful than, the procedural knowledge acquired by the participants in the practice groups; the latter presumably included a goal structure.

A more plausible possibility is that a subset of the incidental training participants discovered that the memorization sequences were patterned, and that they used the pattern as a mnemonic device. Given knowledge of the underlying pattern, one can derive some letters of the sequence from others via inferences that are of the same type as the inferences required to extrapolate a pattern. If some participants engaged in this type of reconstructive recall, they might have acquired some of the same procedural knowledge as the participants in the practice groups. Because each memorization sequence consisted of different letters, they had opportunity and incentive to make this knowledge abstract. In fact, with 18 trials, as opposed to the 3 trials of the practice groups, those incidental training participants who discovered the pattern had more opportunity to construct abstract procedural knowledge than the practice participants. Once again, we see a case of procedural knowledge that was acquired via practice and that transferred from training to problem solving.

GENERAL DISCUSSION

We formulate a set of general learning principles that are consistent with our findings. We also reflect on the limitations and the broader lessons of our work.

Theoretical Implications

The pattern of findings described in the previous section is consistent with the following principles of learning:

Principle 1. Direct instruction via discourse generates declarative knowledge.

Principle 2. Declarative knowledge tends to be expressed at a high level of abstraction and is easy to transfer to a new situation.

Principle 3. Procedural knowledge generated by compiling a declarative structure does not necessarily preserve abstraction level. Instead, knowledge compilation tends to generate task specific procedural knowledge that requires cognitive work to transfer to a new situation.

Principle 4. Practice generates procedural knowledge.

Principle 5. Procedural knowledge generated via practice is of a high level of abstraction and easy to transfer to a new situation.

Principle 6. The transferability of practice-generated procedural knowledge is higher if the practice problems are varied than if they are similar.

Principle 7. Incidental training does generate relevant knowledge, but the learner might not be able to draw upon it to support deliberate problem solving.

Principles 1, 2, 4 and 6 are in accord with prior research and common sense. However, prior research did not provide any reason to expect Principles 3, 5 and 7. Together, they say that there are different sources for procedural knowledge, and that properties such as abstraction level vary with the source. If true, this conjecture has important consequences.

First, consider the idea that the compilation of declarative knowledge does not preserve abstraction level but tends to generate highly specific procedural knowledge. This principle has the potential to explain a large number of negative results in the literature. Our direct instruction scenario was designed to share key features with a typical textbook chapter in a problem solving topic such as physics or statistics: Teach the 'theory' of the topic first, and require independent problem solving afterwards. The expectation is that a good understanding of the theoretical (declarative) part, in combination with problem solving practice, will generate a flexible and general problem solving competence. The failure to realize this expectation is the constant frustration of teachers. The frustration is typically interpreted to mean that the students did not correctly assimilate or understand the declarative part of the instruction; if they had done so, they would know what to do on the problem solving exercises. Numerous studies that document failure to achieve conceptual understanding can be cited in apparent support of this interpretation. The natural response is to search for better ways of teaching for understanding.

That this response is not self-evident is shown by the occasional study that documents poor performance in the present of correct conceptual understanding (Resnick & Omansson, 1987). If our interpretation of our results is accurate, we can understand how those situations arise. The compilation of the declarative understanding does not result in general and applicable procedural knowledge. The radical implication is that this effect is to be expected regardless of the level of conceptual understanding. Improving the students' understanding does not help, because the problem resides in the knowledge compilation process, not in deficiencies in the declarative representation. The compilation process does not preserve abstraction level and so does not support transfer of the procedural knowledge it produces, regardless of depth or completion of the declarative knowledge.

On the other side of the coin, the above principles imply that the advantages of practice are even stronger than we have previously understood. If procedural knowledge created during practice is specific to the practice problems encountered, it is apparently also easy to generalize so as to fit a new situation. The transfer step in our experiments was admittedly short and simple, but it is nevertheless conspicuous how easily it was negotiated even by the low-variability group. In addition, it appears that even the seemingly different task of memorizing letter sequences can generate procedural knowledge that transfers to sequence extrapolation. These results indicate that transfer of procedural knowledge generated via practice is not as limited as one would expect from the traditional view of procedural knowledge as context-bound.

In summary, our results call into question the received view of the relation between declarative and procedural knowledge. The supposed advantages of declarative knowledge as abstract and context-independence are real, but only as long as we are talking about declarative transfer. A person can transfer a pattern readily enough, and hence see a new situation in the light of a previously learned pattern. However, this ability is of limited value and may or may not be expressed in overt behavior, because it does not bring with it an equally powerful and flexible ability to decide what to do in the new situation. The obstacle resides in the process of articulating the action implications of the declarative knowledge, not necessarily in the declarative knowledge itself. In contrast, procedural knowledge generated via practice, supposedly context independent, was here shown to be easier to transfer. Consequently, we have to call into question the usual trade-off explanation for why people have these two types of knowledge (Anderson, 1983; Winograd, 1975). Our findings also call into question the tendency in pedagogical contexts to seek methods to improve students' conceptual understanding for the purpose of, and in the hope of, producing better and more flexible problem solving performance.

Limitations

The present studies have several limitations. The most obvious one is that the training phase was short. The experimental procedure took less than 90 minutes. It is not obvious which effects observed in this time band replicate at the time band in which realistic learning occurs, usually months or even years. We note that academic instruction – which resembles the scenario for direction instruction in experiment 1 – is precisely the context in which it is often claimed that instruction is ineffective, in the sense that students cannot apply what they have learned to solve problems flexibly, so there is at least a thematic similarity between the performance of our instruction groups and classroom observations. Nevertheless, generalizability across time bands is a matter of concern.

A second limitation is the nature of our transfer test. The pattern underlying the transfer problem was a simple transformation of the target pattern. We do not know how the different groups would fare if they attempted a problem based on a completely different pattern. With increased transfer distance, perhaps a relative advantage will appear for the instruction group. There is no evidence for this in our data, but only fresh data can tell.

A third limitation is the fact that our outcome measures were not probing enough. We need to know more about the state of knowledge of the learner at the end of training. The traditional notion of a probing what has been learned with transfer problems is useful but does not go far enough. At the very least, we need to provide multiple transfer problems, each differing from the target problem in a different way. However, we can go further and specify a series of tasks that are not extrapolation problems, but that might involve the same knowledge and reveal how that knowledge is encoded with respect to type and level of abstraction. In short, future studies need to use a multi-dimensional approach to the assessment of what is learned in the alternative scenarios.

Broader View

The observation that cognitive tasks can be mastered in different ways is not as trivial as it might appear at first glance. Direct instruction, practice, and incidental learning differ radically with respect to what information they make available to the learner, how that information is presented, and which cognitive processes are required to make use of it. Direct instruction requires, at the very least, discourse comprehension, but also compilation of the presented information. Because the relevant information is stated explicitly, the learner's attention is directed to it. In a practice scenario, on the other hand, the learner's spontaneous attention allocation might make all the difference; does he or she notice the critical features of the task environment? Practice scenarios present information in the context of the relevant goals, a type of information not necessarily present in a direct instruction scenario. Incidental learning scenarios require that learners automatically encode patterns in the environment even when they are working on a task that does not require them to do so, and that they can retrieve those patterns later for a novel purpose. Learning by observing someone else perform a task differs in yet other ways from all three scenarios explored here. The fact that people can master some tasks via more than one of these scenarios is a phenomenon in its own right.

The methodology we adopted for studying this phenomenon does not follow the standard experimental canon of varying only a single, well-specified, quantitative parameter of the experimental situation in order to be able to make strong inferences about causal relations. Attempting to describe the differences between the three learning scenarios in terms of values on multiple independent variables does not seem a useful exercise. However, if the differences cannot be reduced to parametric differences in any meaningful way, then the dictum that one is to vary one variable at a time excludes comparisons between different learning scenarios from study. But the existence of multiple paths to mastery is a fact, and cognitive research on learning can only loose by refusing to investigate this fact.

In the present studies, we combined qualitatively defined scenarios with parametric variations. We implemented each type of scenario twice, with a parametric difference between the variants (long or short instruction; variable or identical practice problems; 6 versus 18 memorization strings). In general, the results show that the differences between the parametric variants were of smaller magnitude than the differences between the scenarios. The fact that the within-scenario variants replicated the same qualitative pattern in the solution times, while the between-scenario comparisons revealed different patterns of findings increases our confidence that the results indicate theoretically meaningful differences in cognitive processing.

The differences between the scenarios constitute a set of phenomena that a general learning theory ought to be able to account for. That is, a learning theory should be able to predict, for a given path to mastery, what knowledge is acquired along that path and how that knowledge is encoded with respect type, abstraction level, and other properties, and to predict the pattern of differences in a set of outcome measures applied to several paths. It is possible that our particular findings will not replicate in a different task domain, but for every task domain there is some pattern of such differences, and that pattern is a phenomenon against which we can test any learning theory that claims generality.

In particular, a computational model that claims generality must be able to simulate the attainment of mastery of a given task in any scenario in which people can master that task. This is a *sufficiency criterion* (Newell & Simon, 1972) that current models of learning do not pass. For example, a model of learning from instruction might be build on top of the sophisticated models of discourse comprehension proposed by Kintsch (1998) and others. Although those models are informative, they do not explain what information a learner extracts from practice. This is not a criticism of comprehension models as models of comprehension. Conversely, computational models of learning by doing have been very successful in modeling practice effects such as the power law of learning (Anderson, 1982), but the learning mechanisms they postulate cannot explain how a person might learn from instruction (but see Ohlsson, Ernst, & Rees, 1991 for an attempt in this direction). Again, this is not a critique of such theories as theories of practice effects, only a statement about a boundary of application. Theories of incidental and implicit learning have focused on specifying the inductive processes that are presumably operating during training, but those processes do not explain how a learner benefits from either instruction or deliberate practice. To highlight the boundaries of the application of these theories is not to critique them for the purpose for which they were designed. Our point is that neither theories of discourse comprehension, practice or incidental induction are candidates for general theories of learning, because people can learn from discourse *and* practice *and* incidental induction *and* in yet other ways.

This situation should not be viewed as problematic. Science progresses by building local theories for particular domains and then attempting to unify them under more fundamental principles. The fact that local theories of learning have reached a certain level of theoretical sophistication holds out the promise that unification has arrived at the threshold of the possible. Unified theories of learning must be tested against phenomena with wider scope than the effects of parametric variations within narrowly defined learning scenarios. The pattern of cross-scenario comparisons reported in this article constitute one example of such a phenomenon.

ACKNOWLEDGEMENT

We thank Brandy Jones and Chris Stanely for help in collecting and analyzing the data.

REFERENCES

Altmann, G. T. M., Dienes, A., & Goode, A. (1995). Modality independence of implicitly learned grammatical knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 899-912.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*, 369-406.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Berry, D., (Ed.). (1997). *How implicit is implicit learning?* Oxford, UK: Oxford University Press.

Cohen, J. (1988). *Statistical power analysis of the behavioral sciences*. (2nd Ed.) New York: Academic Press.

Cormier, S. M., & Hagman, J. D., (Eds.), (1987). *Transfer of learning*. San Diego, CA: Academic Press.

Detterman, D. K., & Sternberg, R. J., (Eds.), (1993). *Transfer on trial*. Norwood, NJ: Ablex.

Dienes, A. P. (1963). *An experimental study of mathematics learning*. London, UK: Hutchinson.

Friedman, M. (1978). The manipulative materials strategy: The latest pied piper? *Journal for Research in Mathematics Education*, *9*, 78-80.

Greeno, J. G., & Simon, H. A. (1974). Processes for sequence production. *Psychological Review, 81*, 187-198.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. London, UK: Cambridge University Press.

Klahr, D., & Wallace, J. G. (1970). The development of serial completion strategies: An information-processing analysis. *British Journal of Psychology, 61*, 243-257.

Kotovsky, K., & Simon, H. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology, 4*, 399-424.

Manza, L. & Reber, A. S. (1997). Representing artificial grammars: Transfer across stimulus forms and modalities. In D. C. Berry (Ed.), *How implicit is implicit learning?* (pp. 73-106). Oxford, UK: Oxford University Press.

McKeough, A., Lupart, J., & Marini, A., (Eds.), (1995). *Teaching for transfer*. Mahwah, NJ: Erlbaum.

Neves, D. M., & Anderson, J. R. (1981). Knowledge compilation: Mechanisms for the automatization of cognitive skills. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 57-84). Hillsdale, NJ: Erlbaum.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nokes, T. J., & Ohlsson, S. (2000). An inquiry into the function of implicit knowledge and its role in problem solving. In L. R. Gleitman and A. K. Joshi, (Eds.), *Proceedings of the Twenty Second Annual meeting of the Cognitive Science Society* (pp. 829-834), Mahaw, NJ: Lawrence Erlbaum.

Ohlsson, S. (1993). Abstract schemas. *Educational Psychologist, 28*, 51-66.

Ohlsson, S. (1994). Declarative and procedural knowledge. In T. Husen & T. Neville-Postlethwaite, (Eds.), *The International Encyclopedia of Education* (Vol. 3, 2nd ed., pp.1432-1434). London, UK: Pergamon Press.

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

Ohlsson, S., Ernst, A., & Rees, E. (1992) The cognitive complexity of doing and learning arithmetic. *Journal for Research in Mathematics Education, 23*, 441-467.

Ohlsson, S., & Lehtinen, E. (1997). Abstraction and the acquisition of complex ideas. *International Journal of Educational Research, 27*, 37-48.

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

Olejnik, S. & Algina, J. (2000). Measures for effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*, 241-286.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher, 6*, 16-25.

Perruchet, P. & Gallego, J. (1997) A subjective unit formation account of implicit learning. In D. C. Berry, (Ed.), *How implicit is implicit learning*? (pp. 124-161). Oxford, UK: Oxford University Press.

Postman, L. (1964). Short-term memory and incidental learning. In A. W. Melton, (Ed.), *Categories of Human Learning* (pp. 145-201). New York, NY: Academic Press Inc.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*, 317-327.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118*, 219-235.

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. New York: Oxford University Press.

Resnick, L. B., & Omanson, S. F. (1987). Learning to understand arithmetic. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 3, pp. 41-95). Hillsdale, NJ: Erlbaum.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review, 77*, 481-495.

Restle, F., & Brown, E. (1970). Organization of serial pattern learning. In G. Bower, (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 4). New York: Academic Press.

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology, 39*, 475-543.

Royer, J. M. (1979). Theories of the transfer of learning. *Educational Psychologist, 14*, 53-69.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomena. *Educational Psychologist, 24*, 113-142.

Seger, C. A. (1994). Implicit Learning. *Psychological Bulletin, 115*, 163-196.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 501-518.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill.* Cambridge, MA: Harvard University Press.

Simon, H. (1972). Complexity and the representation of patterned sequences of symbols. *Psychological Review, 79*, 369-382.

Simon, H. A., & Kotovsky K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review, 70*, 534–546.

Skemp, R. R. (1971). *The psychology of learning mathematics*. Middlesex, UK: Penguin.

Sowell, E. J. (1989). Effects of manipulative materials in mathematics instruction. *Journal for Research in Mathematics Education, 20*, 498-505.

Squire, L. (1987). *Memory and brain.* New York: Oxford University Press.

Stadler, M., & Frensch, P, (Eds.), (1998). *Handbook of implicit learning*. Thousand Oaks, CA: SAGE.

Thurstone, L. & Thurstone, T. (1941). *Factorial studies in intelligence.* Chicago: University of Chicago Press.

Wattenmaker, W. D. (1999). The influence of prior knowledge in intentional versus incidental concept learning. *Memory & Cognition, 27*, 685-698.

Winograd, T. (1975). Frame representations and the declarative-procedural controversy. In D. G. Bobrow and A. Collins, (Eds.), *Representation and understanding* (pp.185-210). New York: Academic Press.

Appendix A

Incidental Learning Materials

Problem 1

1)	W	X	V	Y	U	G	X	Y	W	Z	V	H
2)	F	G	E	H	D	P	G	H	F	I	E	Q
3)	K	L	J	M	I	U	L	M	K	N	J	V
4)	R	S	Q	T	P	B	S	T	R	U	Q	C
5)	M	N	L	O	K	W	N	O	M	P	L	X
6)	I	J	H	K	G	S	J	K	I	L	H	T
7)	C	D	B	E	A	M	D	E	C	F	B	N
8)	Q	R	P	S	O	A	R	S	Q	T	P	B
9)	L	M	K	N	J	V	M	N	L	O	K	W
10)	U	V	T	W	S	E	V	W	U	X	T	F
11)	N	O	M	P	L	X	O	P	N	Q	M	Y
12)	V	W	U	X	T	F	W	X	V	Y	U	G
13)	T	U	S	V	R	D	U	V	T	W	S	E
14)	J	K	I	L	H	T	K	L	J	M	I	U
15)	O	P	N	Q	M	Y	P	Q	O	R	N	Z
16)	S	T	R	U	Q	C	T	U	S	V	R	D
17)	D	E	C	F	B	N	E	F	D	G	C	O
18)	H	I	G	J	F	R	I	J	H	K	G	S

Problem 2

1) B D Y E C X X E G W H F V V

2) G I T J H S S J L R M K Q Q

3) T V G W U F F W Y E Z X D D

4) N P M Q O L L Q S K T R J J

5) D F W G E V V G I U J H T T

6) R T I U S H H U W G X V F F

7) J L Q M K P P M O O P N N N

8) P R K S Q J J S U I V T H H

9) C E X F D W W F H V I G U U

10) F H U I G T T I K S L J R R

11) O Q L R P K K R T J U S I I

12) K M P N L O O N P N Q O M M

13) S U H V T G G V X F Y W E E

14) E G V H F U U H J T K I S S

15) Q S J T R I I T V H W U G G

16) H J S K I R R K M Q N L P P

17) M O N P N M M P R L S Q K K

18) I K R L J Q Q L N P O M O O

Table Captions

Table 1. Two sequence extrapolation problems and their associated transfer problems.

Table 2. Three practice problems for each problem type.

Table 3. Mean problem solving accuracy scores and standard deviations for the practice, direct instruction, and no-training groups on target and transfer problems.

Table 4. Mean solution times (seconds) and standard deviations for the practice, direct instruction and no-training groups on target and transfer problems.

Table 5. Mean problem solving accuracy scores and standard deviations for the practice, implicit learning, and no-training groups on target and transfer problems.

Table 6. Mean solution times (seconds) and standard deviations for the practice, implicit learning and no-training groups on target and transfer problems.

Table 7. Measures of transfer based on accuracy performance for all training groups on both target and transfer problems.

Table 8. Deliberation and extrapolation times for the control group and each training group in experiments 2 and 3.

Table 9. Measures of transfer based on solution times for all training groups on both target and transfer problems.

Table 10. Mean solution times and standard deviations for the high variability practice and long direct instruction groups on target and transfer problems.

Table 1

Two Sequence Extrapolation Problems and Their Associated Transfer Problems

Problem-Type	Given letter/ number sequence	Correct 8-step extrapolation
Problem 1		
Target	E F D G C O F G E H D P	G H F I E Q H I
Transfer	E G D I C O G I F K E P	I K H M G Q K M
Problem 2		
Target	A C Z D B Y Y D F X G E W W	G I V J H U U J
Transfer	A E Z G C X X G K V M I T T	M Q R S O P P S

Table 2

Three Practice Problems for each Problem-Type

Problem-Type	Given letter/ number sequence	Correct 8-step extrapolation
Problem 1		
1	I J H K G S J K I L H T	K L J M I U L M
2	R S Q T P B S T R U Q C	T U S V R D U V
3	M N L O K W N O M P L X	O P N Q M Y P Q
Problem 2		
1	G I T J H S S J L R M K Q Q	M O P P N O O P
2	R T I U S H H U W G X V F F	X Z E A Y D D U
3	N P M Q O L L Q S K T R J J	T V I W U H H W

Table 3

Mean Problem Solving Accuracy Scores and Standard Deviations for the Practice, Direct Instruction, and No-training Groups on Target and Transfer Problems

	Target Problem		Transfer Problem	
Training Group	All Participants	*Upper 2/3	All Participants	*Upper 2/3
No-training	3.58 (2.86)	---	3.10 (2.84)	---
Short Instruction (3.07)	4.04 (3.01)	4.37 (2.92)	3.23 (2.92)	3.71
Long Instruction (2.21)	5.28 (2.63)	6.10 (2.56)	4.88 (2.69)	6.08
Practice Low (2.10)	6.32 (2.47)	7.25 (1.63)	5.27 (2.85)	6.38
Practice High (1.72)	5.65 (3.01)	7.33 (1.10)	4.94 (2.95)	6.52

**Note*. Upper 2/3 refers to the top two-thirds of participants who were most successful on the training task.

Table 4

Mean Solution Times (seconds) and Standard Deviations for the Practice, Direct Instruction and No-training Groups on Target and Transfer Problems

	Target Problem		Transfer Problem	
Training Group	All Participants	*Upper 2/3	All Participants	*Upper 2/3
No-training	247.2 (82.2)	---	252.0 (73.8)	---
Short Instruction	231.0 (76.2)	219.6 (66.0)	220.8 (76.8)	211.2
(72.6)				
Long Instruction	223.8 (70.8)	210.1 (69.6)	262.8 (72.0)	259.2
(63.6)				
Practice Low	144.6 (83.4)	131.4 (75.6)	175.8 (88.2)	162.6
(79.8)				
Practice High	106.8 (50.4)	91.2 (39.0)	152.4 (64.8)	150.6
(48.0)				

**Note*. Upper 2/3 refers to the top two-thirds of participants who were most successful on the training task.

Table 5

Mean Problem Solving Accuracy Scores and Standard Deviations for the Practice, Incidental Learning, and No-training Groups on Target and Transfer Problems

	Target Problem		Transfer Problem	
Training Group	All Participants	*Upper 2/3	All Participants	*Upper 2/3
No-training	3.58 (2.86)	---	3.10 (2.84)	---
Incidental Short (2.59)	3.34 (2.67)	3.09 (2.95)	2.56 (2.68)	2.59
Incidental Long (2.40)	3.87 (2.58)	4.22 (2.26)	3.33 (2.53)	3.75
Practice Low (2.02)	5.56 (3.13)	7.03 (1.89)	4.60 (2.59)	5.69
Practice High (2.39)	5.54 (2.64)	6.69 (2.10)	4.85 (2.67)	5.90

**Note*. Upper 2/3 refers to the top two-thirds of participants who were most successful on the training task.

Table 6

Mean Solution Times (seconds) and Standard Deviations for the Practice, Incidental Learning and No-training Groups on Target and Transfer Problems

	Target Problem		Transfer Problem	
Training Group	All Participants	*Upper 2/3	All Participants	*Upper 2/3
No-training	247.2 (82.2)	---	252.0 (73.8)	---
Incidental Short (109.8)	217.7 (89.6)	209.1 (103.3)	230.5 (96.0)	218.5
Incidental Long (94.0)	191.9 (66.4)	194.9 (66.7)	213.2 (92.3)	223.7
Practice Low (65.8)	129.7 (43.8)	118.1 (41.1)	161.9 (61.2)	161.7
Practice High (63.1)	141.5 (88.5)	118.9 (64.4)	188.4 (76.5)	183.0

**Note*. Upper 2/3 refers to the top two-thirds of participants who were most successful on the training task.

Table 7

Measures of Transfer based on Accuracy for all Training Groups on both Target and Transfer Problems

Concept 1	Formula 1		
transfer improvement normalized by control performance	$T_{\%\ \text{Improvement}} = (E - C / C) \times 100$		
Group	All Participants		Upper 2/3
	Target	Transfer	Target
Transfer			
Incidental Short 16.5%	-6.7%	-14.5%	-13.8%
Incidental Long 21.0%	8.1%	7.1%	17.7%
Practice Low 95.2%	66.4%	60.0%	99.3%
Practice High 101.8%	56.4%	58.0%	96.9%
Short Instruction 19.7%	12.8%	4.2%	21.9%
Long Instruction 96.0%	47.5%	57.4%	70.2%

Cont'd

Table 7, Cont'd

Concept 2	Formula 2
Percent of transfer improvement of total learning possible	$T_{\%\ Total\ Learning} = (E - C /_{Perform\ limit} - C) \times 100$

Group	All Participants		Upper 2/3
	Target	Transfer	Target
Transfer			
Incidental Short	-5.5%	-11.0%	-11.2% -
10.4%			
Incidental Long	6.5%	4.6%	14.4%
13.3%			
Practice Low	53.9%	37.9%	80.6%
60.3%			
Practice High	45.7%	36.7%	78.6%
64.4%			
Short Instruction	10.2%	2.7%	17.8%
12.5%			
Long Instruction	38.5%	36.4%	57.0%
60.7%			

Notes: Upper 2/3 refers to the top two-thirds of participants who were most successful on the training task.

Table 8

Deliberation and Extrapolation Times for the Control Group and Each Training Group in Experiments 2 and 3

Experiment 2: Upper Two Thirds Analysis, *n* = ~20

	Deliberation Time		*Extrapolation Time*	
Training Group Transfer (SD)	Target (SD)	Transfer (SD)	Target (SD)	
No-training (57)	122 (61)	126 (73)	125 (63)	126
Short Instruction (57)	60 (35)	50 (33)	159 (58)	161
Long Instruction (56)	54 (35)	45 (28)	156 (55)	214
Practice Low 96 (43)	56 (47)	66 (53)	75 (44)	
Practice High 102 (38)	32 (18)	48 (22)	59 (27)	

Cont'd

Table 8, Cont'd

Experiment 3: Upper Two Thirds Analysis, *n* = ~17

	Deliberation Time		*Extrapolation Time*	
Training Group	Target (SD)	Transfer (SD)	Target (SD)	
Transfer (SD)				
No-training	122 (61)	126 (73)	125 (63)	126
(57)				
Incidental Short	107 (56)	109 (72)	102 (66)	109
(76)				
Incidental Long	97 (46)	85 (51)	98 (47)	
138 (79)				
Practice Low	47 (29)	48 (31)	71 (24)	114
(52)				
Practice High	44 (43)	66 (59)	75 (36)	117
(45)				

Table 9

Measures of Transfer based on Solution Times for all Training Groups on both Target and Transfer Problems

Concept 1	Formula 1		
transfer improvement normalized by control performance	$T_{\% \text{ Improvement}} = (C - E / C) \times 100$		
Group	All Participants		Upper 2/3
	Target	Transfer	Target
Transfer			
Incidental Short	11.9%	8.5%	15.4%
13.2%			
Incidental Long	22.3 %	15.3%	21.0%
11.2%			
Practice Low	44.2%	32.8%	49.3%
35.6%			
Practice High	50.7%	33.4%	58.2%
34.7%			
Short Instruction	6.6%	12.6%	11.2%
16.2%			
Long Instruction	9.4%	-4.3%	14.7%
3.0%			

Cont'd

Table 9, Cont'd

Concept 2	Formula 2			
Percent of transfer improvement of total learning possible	$T_{\%\ Total\ Learning} = (C - E / C - _{Perform\ limit}) \text{ x } 100$			
Group	All Participants		Upper 2/3	
	Target	Transfer	Target	Transfer
Incidental Short	14.3%	12.5%	18.5%	19.6%
Incidental Long	26.9%	22.7%	25.4%	16.5%
Practice Low	53.3%	48.5%	59.4%	52.7%
Practice High	61.1%	49.5%	70.2%	51.3%
Short Instruction	8.0%	18.3%	13.5%	24.0%
Long Instruction	11.3%	-6.4%	17.8%	4.4%

Notes: Upper 2/3 refers to the top two-thirds of participants who were most successful on the training task.

Table 10

Mean Solution Times (seconds) and Standard Deviations for the High Variability Practice and Long Direct Instruction Groups on Target and Transfer Problems

	Deliberation Time		*Extrapolation Time*	
Training Group	Target (SD)	Transfer (SD)	Target (SD)	Transfer (SD)
High-level Accuracy (8 correct extrapolations)				
Long Instruction (*n* = 9)	43 (28)	42 (24)	128 (34)	211 (59)
Practice High (*n* = 14)	34 (29)	53 (34)	46 (32)	94 (33)
Mid-level Accuracy (5-7 correct extrapolations)				
Long Instruction (*n* = 8)	51 (22)	60 (34)	174 (79)	227 (52)
Practice High (*n* = 8)	50 (24)	64 (39)	71 (28)	117 (42)
Low-level Accuracy (0-4 correct extrapolations)				
Long Instruction (*n* = 13)	82 (12)	62 (38)	178 (66)	191 (67)
Practice High (*n* = 9)	52 (34)	61 (63)	85 (45)	73 (63)

Figure Captions

Figure 1. Example target and transfer problem with the relations identified.

Figure 2. Example problem as it was presented on the computer.

Figure 3. Mean problem solving scores for the no-training group on target and transfer problems for both problem types.

Figure 4. Mean solving time scores for the no-training group on target and transfer problems for both problem types.

Figure 5. Diagrammatic pattern illustration.

Figure 6. Mean problem solving scores for the low and high variability practice groups on the three training problems.

Figure 7. The proportion of correctly recalled pattern relations for short and long instruction groups for both problem types.

Figure 8. Mean problem solving scores for the low and high practice groups on the three training problems.

Figure 9. Mean memory scores for both low and high incidental training groups on patterns 1 and 2.

Figure 10. The proportion of specific pattern relations and general concepts identified by incidental groups for each pattern.

Figure 1

Problem Type Identified Relations

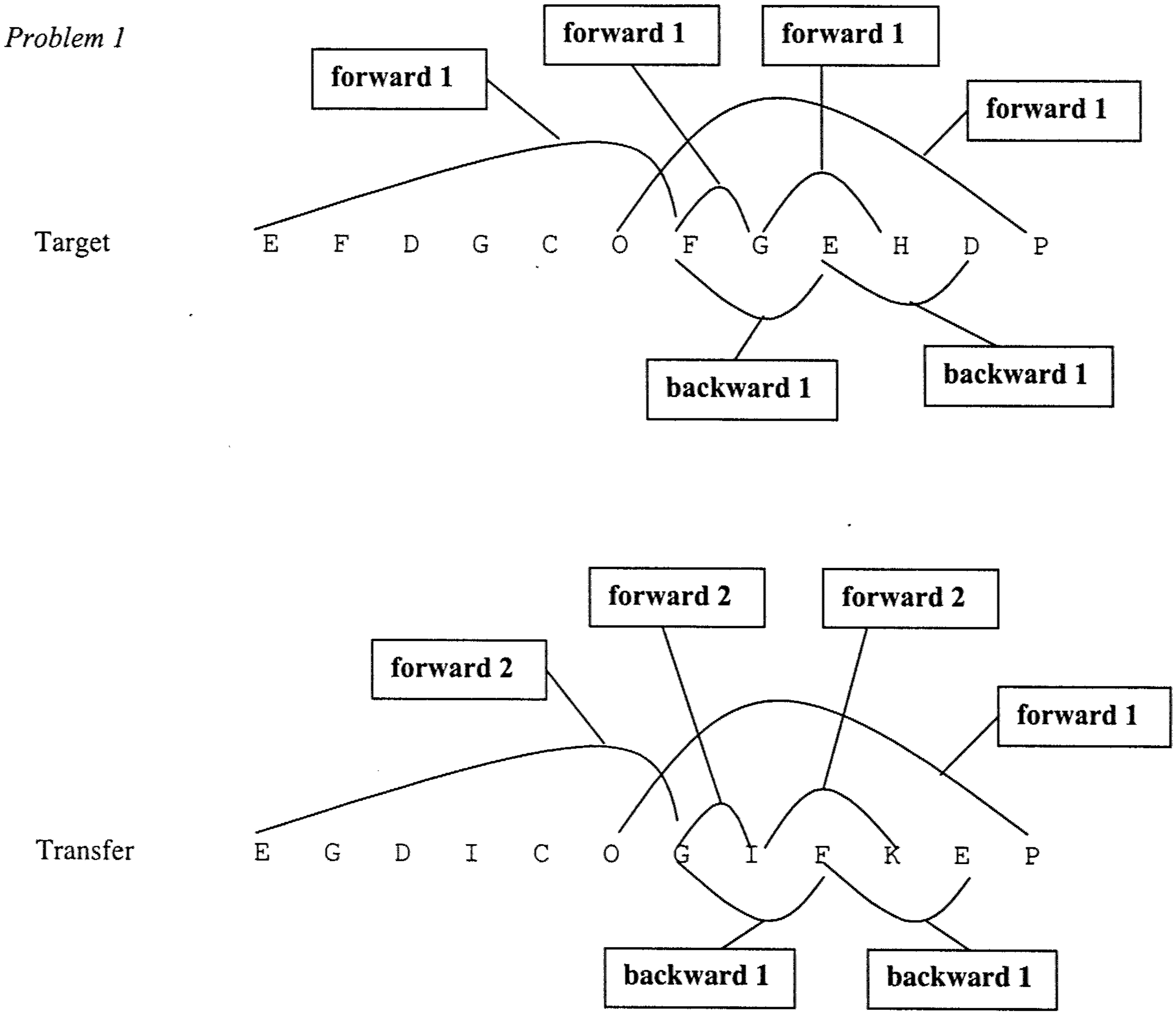


Figure 2

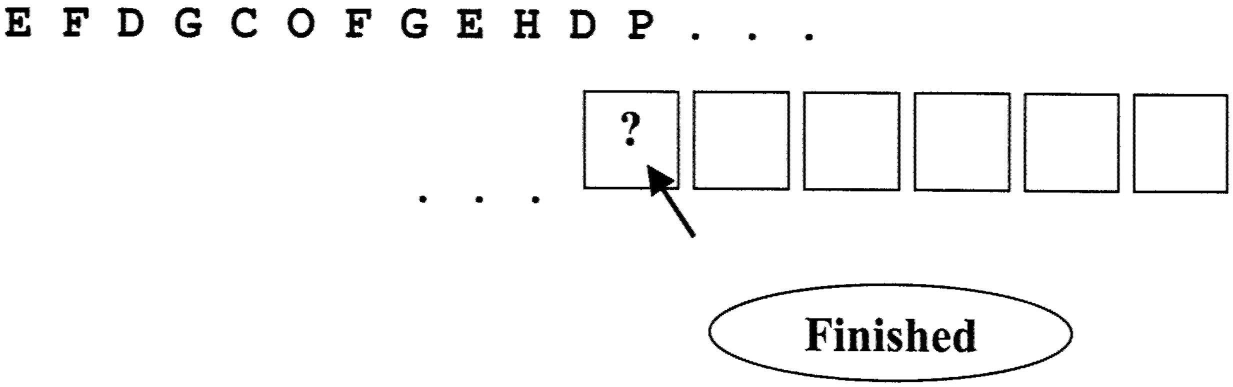


Figure 3

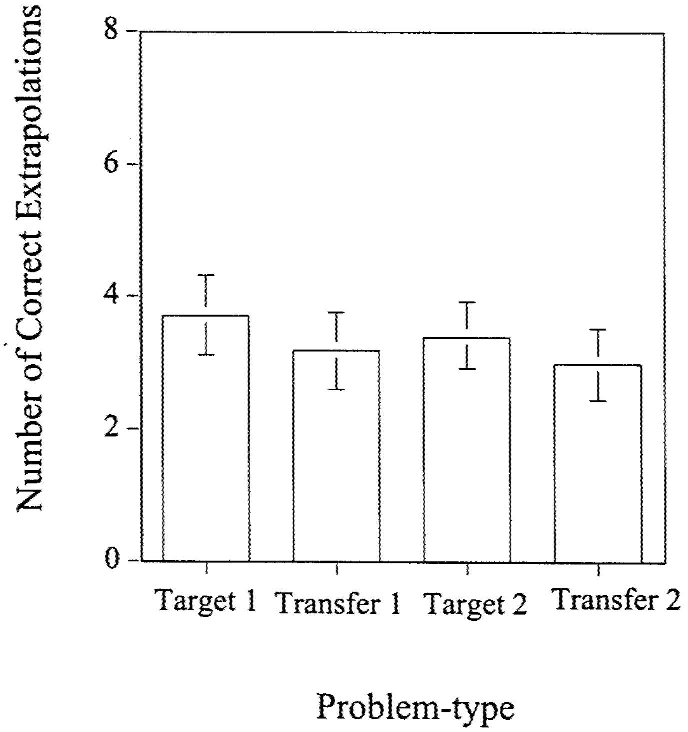

Number of Correct Extrapolations
8
6
4
2
0
Target 1
Transfer 1
Target 2
Transfer 2
Problem-type

Figure 4

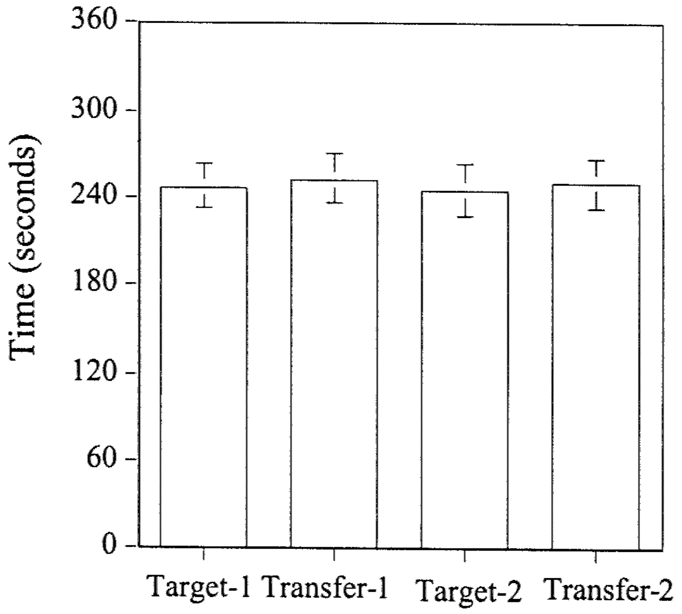
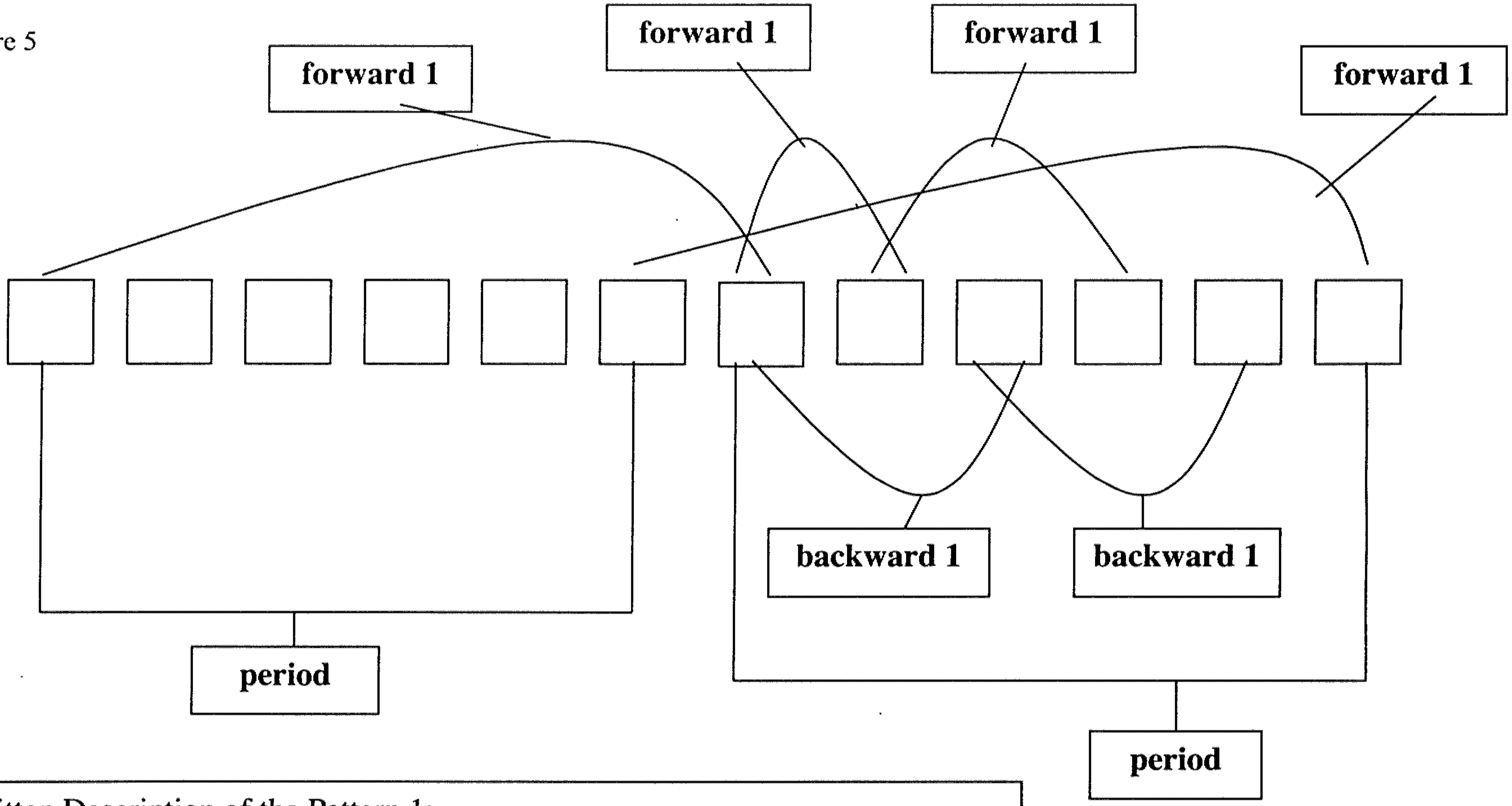

Time (seconds)
360
300
240
180
120
60
0
Target-1
Transfer-1
Target-2
Transfer-2

Figure 5



Written Description of the Pattern 1:

The period is 6 letters long. The first letter in the period is one step forward from the first letter from previous period. The second letter in the period is one step forward from the first letter. The third letter is one step backwards from the first letter. The fourth letter is one step forward from the second letter. The fifth letter is one step backward from the third letter. And the last letter of the period is one step forward from the last letter from the previous period.

Figure 6

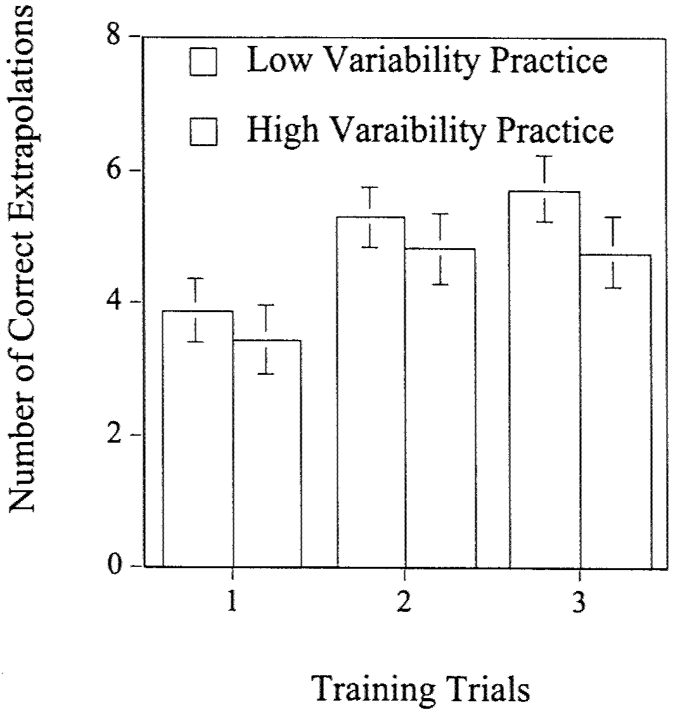

Number of Correct Extrapolations
8
6
4
2
0
□ Low Variability Practice
□ High Varaibility Practice
1
2
3
Training Trials

Figure 7

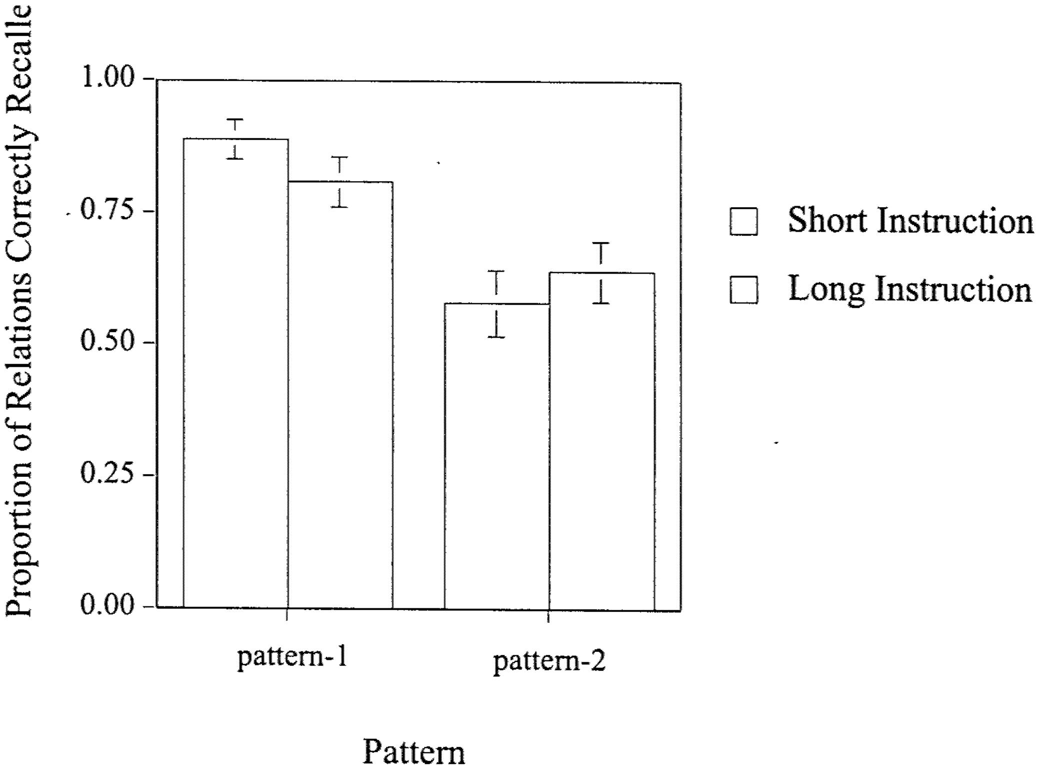

Proportion of Relations Correctly Recalle
1.00
0.75
0.50
0.25
0.00
pattern-1
pattern-2
Pattern
□ Short Instruction
□ Long Instruction

Figure 8

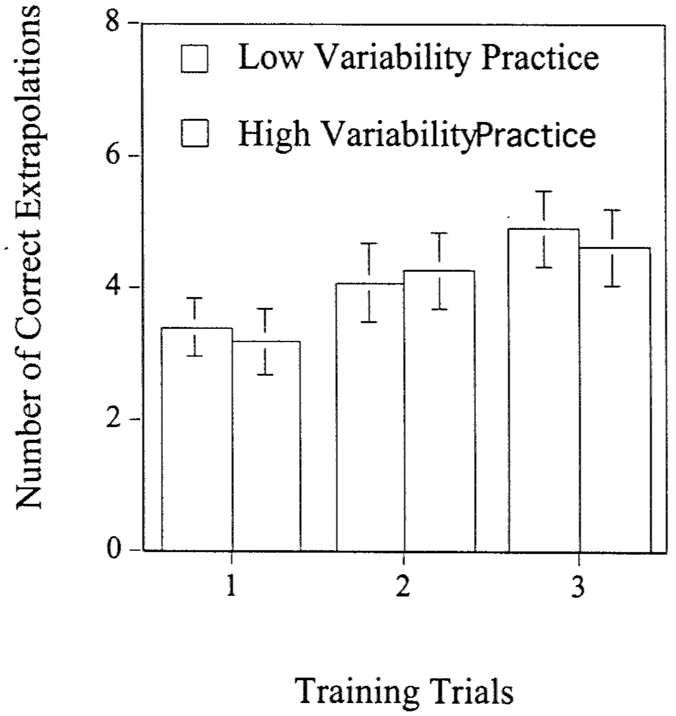
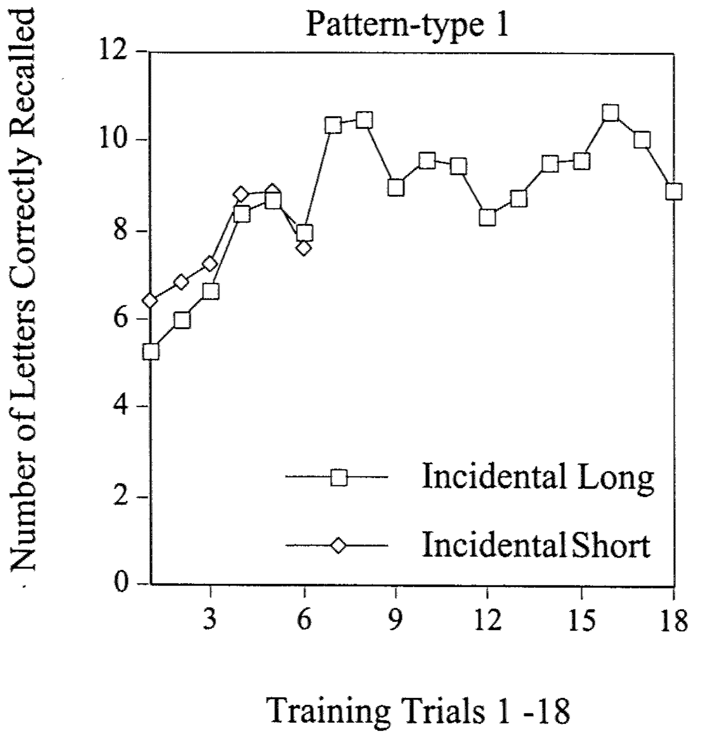

Number of Correct Extrapolations
8
6
4
2
0
□ Low Variability Practice
□ High VariabilityPractice
1
2
3
Training Trials

Figure 9


Pattern-type 1
Number of Letters Correctly Recalled
12
10
8
6
4
2
0
3
6
9
12
15
18
Incidental Long
Incidental Short
Training Trials 1 -18

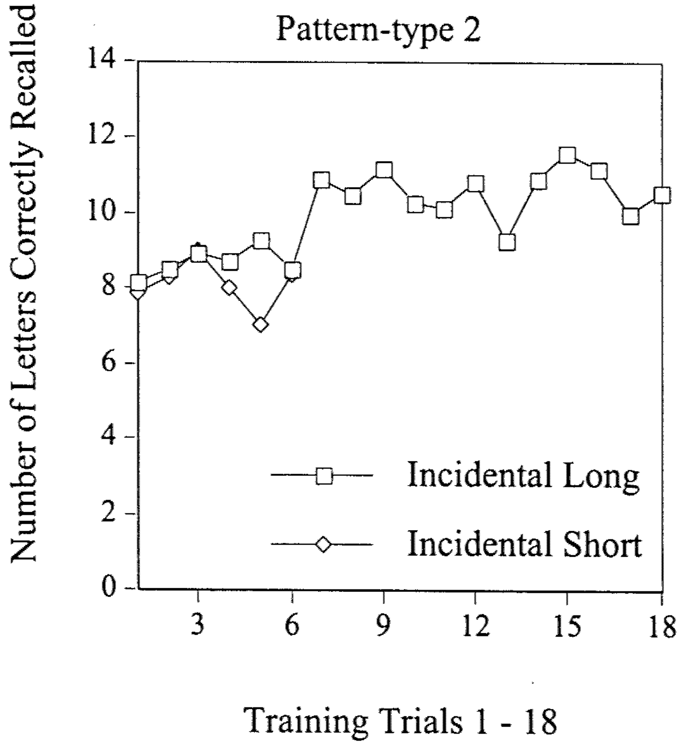
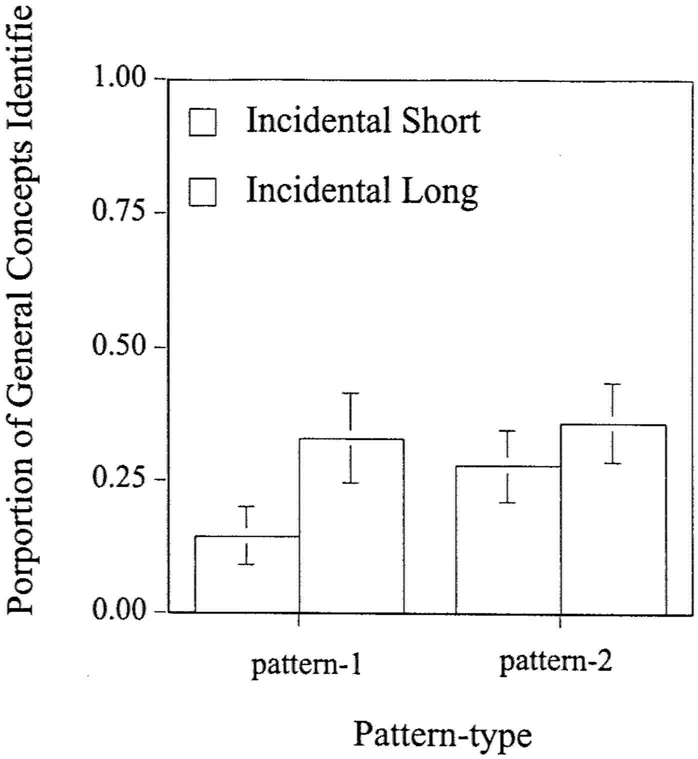
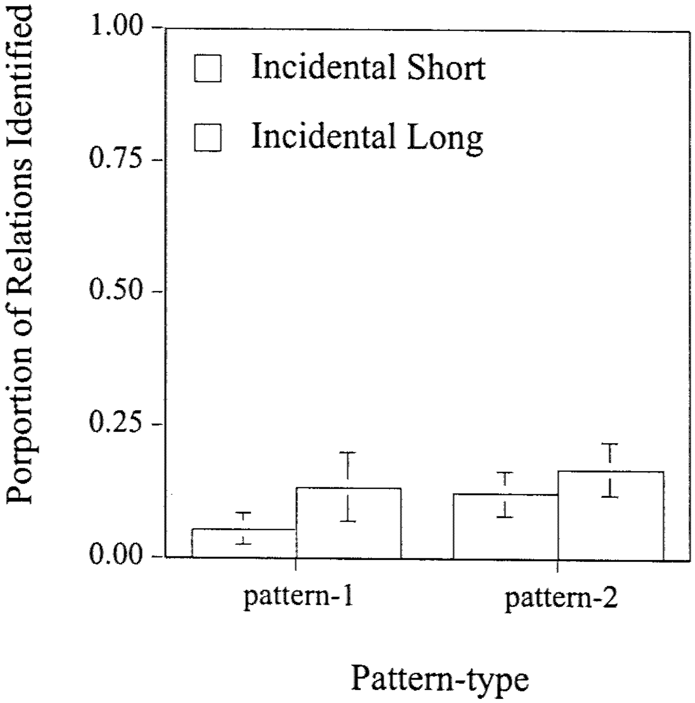

Pattern-type 2
Number of Letters Correctly Recalled
14
12
10
8
6
4
2
0
3
6
9
12
15
18
Incidental Long
Incidental Short
Training Trials 1 - 18

Figure 10


Porportion of General Concepts Identifie
1.00
0.75
0.50
0.25
0.00
□ Incidental Short
□ Incidental Long
pattern-1
pattern-2
Pattern-type


Porportion of Relations Identified
1.00
0.75
0.50
0.25
0.00
□ Incidental Short
□ Incidental Long
pattern-1
pattern-2
Pattern-type