Journal of Experimental Psychology: General
1991, Vol. 120, No. 1, 34-45

0096-3445/91/$3.00

Immediate and Delayed Transfer of Training Effects in Statistical Reasoning

Geoffrey T. Fong
University of Waterloo
Waterloo, Ontario, Canada

Richard E. Nisbett
University of Michigan

Ss were trained on the law of large numbers in a given domain through the use of example problems. They were then tested either on that domain or on another domain either immediately or after a 2-week delay. Strong domain independence was found when testing was immediate. This transfer of training was not due simply to Ss' ability to draw direct analogies between problems in the trained domain and in the untrained domain. After the 2-week delay, it was found that (a) there was no decline in performance in the trained domain and (b) although there was a significant decline in performance in the untrained domain, performance was still better than for control Ss. Memory measures suggest that the retention of training effects is due to memory for the rule system rather than to memory for the specific details of the example problems, contrary to what would be expected if Ss were using direct analogies to solve the test problems.

Two central questions about the nature of reasoning have been addressed since Plato's time: At what level of generality and abstraction do rules for reasoning exist? and Is it possible to improve people's reasoning abilities? These two questions are intimately related. Plato, and most subsequent thinkers up until the late nineteenth century, believed that people possess very abstract rules and that, as a consequence, it is relatively easy to improve reasoning. According to this perspective, known as *formal discipline*, rules can be taught in the abstract, in the form of mathematics or logic, and they will then be applied across the concrete domains of everyday life. Psychologists and educators through the nineteenth century explained the process by drawing an analogy between physical and mental training: Just as one could exercise the muscles to obtain stronger muscles, so too could one exercise the reasoning faculties to strengthen them.

Twentieth-century psychology has been much less sanguine about the effects of training, partly because there has been substantial resistance to believing that inferential rules exist at a very general or abstract level. In his early studies, which suggested that practice in memorizing poetry did not serve to improve the faculty of memory more generally, William James (1890) called the analogy into question. So, too, did the extensive research program of Thorndike (1906; Thorndike & Woodworth, 1901). Thorndike concluded, on the basis of a number of experiments, that transfer of training was a will-o'-the-wisp that was dependent entirely on whether the target task shared "identical elements" with the task on which subjects had been trained. Thorndike's position was thus characterized by extreme concreteness and domain specificity of training, as is conveyed by the following quotation:

> Training the mind means the development of thousands of particular independent capacities, the formation of countless particular habits, for the working of any mental capacity depends upon the concrete data with which it works. Improvement of any one mental function or activity will improve others only insofar as they possess elements common to it also. The amount of identical elements in different mental functions and the amount of general influence from special training are much less than common opinion supposes. (Thorndike, 1906, p. 246)

Thorndike's (1906) view finds its counterpart today in the positions of such theorists as D'Andrade (1982), Griggs and Cox (1982), Manktelow and Evans (1979), and Reich and Ruth (1982). These theorists hold that deductive reasoning occurs not by virtue of the application of abstract rules of logic but by virtue of local, concrete rules tied to the domain in question.

In contrast to Thorndike's (1906) antiformalist position, Piaget (e.g., Inhelder & Piaget, 1958) and modern theorists in the Piagetian tradition (e.g., Braine, 1978; Braine, Reiser, & Rumain, 1984) hold that extreme concreteness is characteristic only of the young child. By the beginning of adolescence, when the child reaches the stage of formal operations, reasoning is governed by the use of abstract inferential rules that are essentially identical to formal statistical and logical rules. Even Piaget and his followers, however, have been pessimistic about whether these abstract rules can be improved through instruction. Piaget believed that the acquisition of abstract inferential rules was almost entirely dependent on spontaneous cognitive development resulting from active self-discovery and that formal instruction could not accelerate the process to any great extent.

This research was supported by a National Science and Engineering Research Council of Canada and by a Northwestern University grant to Geoffrey T. Fong and by National Science Foundation Grant SES-8507342, National Institute of Mental Health Grant IR01MH38466, and Office of Naval Research Grant 442PT/85-228 to Richard E. Nisbett.

We thank Steven Dukeshire, Albert Erlebacher, Craig Kakuda, Denise Meisel, Joanna Mendel, Jon Soble, and Sandra Yu for their assistance. Patricia Cheng, Carl Duncan, James Hall, Reid Hastie, Win Hill, Keith Holyoak, Ziva Kunda, and Darrin Lehman provided valuable comments on previous drafts of this article.

Correspondence concerning this article should be addressed to Geoffrey T. Fong, Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. Electronic mail may be sent to gfong@watdcs.uwaterloo.ca.

The possibility that reasoning ability might be improved through instruction has thus been met with considerable pessimism throughout the history of modern psychology. It should be noted that although Thorndike's research on transfer of training provided some of the early pessimism for the effects of instruction on reasoning, the studies actually had little to do with reasoning as it would be defined today. Instead, they examined transfer from tasks such as canceling parts of speech and estimation of areas of rectangles. Nevertheless, more recent research on transfer of training for problem-solving tasks has hardly been more encouraging. Strong positive transfer effects seem to be difficult to obtain (see Gick & Holyoak, 1987, for a review of the literature on transfer effects). For example, exposure to the Tower-of-Hanoi problem does not readily transfer to other, formally identical problems (Hayes & Simon, 1977).

Other researchers have attempted to improve intelligence, critical thinking skills, and other skills through formal training (see Nickerson, Perkins, & Smith, 1985, and Resnick, 1987, for reviews). Some of these attempts have been shown to be effective in improving higher order reasoning skills, at least within the classroom. Whether such training generalizes to contexts outside the classroom, however, has not been rigorously examined.

In contrast, a recent set of experiments conducted by Fong, Krantz, and Nisbett (1986) showed that training in statistics strongly influences the way people reason about events involving uncertainty in everyday life and that such training readily transfers to domains outside of the domain of training. For example, Fong et al. presented subjects who had varying degrees of statistical training with a problem about a manufacturer's representative who travels a great deal and tries to maximize the quality of her dining experiences by returning to restaurants where she had an excellent meal on her first visit. She finds, however, that subsequent meals are rarely as good as the first. Subjects were asked to explain why this occurs. Subjects without training in statistics almost invariably gave a purely "deterministic" answer that stressed possible causal explanations. For example, they suggested that "the chefs may change a lot" or "her expectations were so high that she could only be disappointed." Subjects with some training in statistics were more likely to give "statistical" answers, that is, those that made at least some mention of the variability of meal quality at a restaurant over time. These statistically sophisticated subjects were more likely to state that "maybe it was just by chance that she got such a good meal the first time." Finally, subjects with substantial training in statistics were quite likely not only to refer to statistical considerations but also to structure the problem as one involving sample values and population parameters. For example, "there are probably more restaurants where you might get an occasional excellent meal than restaurants that serve only excellent meals; chances are that when she got such a good meal the first time, it was just because she happened to hit it lucky at an inconsistent restaurant."

Fong et al. (1986) argued that statistical training is effective because people possess rudimentary but abstract intuitive versions of the law of large numbers and other statistical principles, or *statistical heuristics* (Nisbett, Krantz, Jepson, & Kunda, 1983). These rules exist in people's cognitive repertoire as the statistical counterpart to the nonstatistical heuristics, such as representativeness and availability, which have been identified by Kahneman and Tversky (1972, 1973; Tversky & Kahneman, 1974). Because people possess some statistical intuitions, it is possible to improve the rule system by relatively formal training procedures that work directly on the abstract rules themselves.

In support of the view that statistical heuristics exist in abstract form rather than simply in the form of concrete, domain-specific rules, Fong et al. (1986) presented two studies using two different training procedures. In the *rule training* condition, subjects were taught about the formal properties of the law of large numbers in a brief training session. The session began with formal definitions of sample, population, and sampling, and ended with a demonstration of the law of large numbers involving the classic gumball-urn model. This formal training increased both the frequency and the quality of statistical reasoning about a wide variety of everyday life problems, from probabilistic problems involving randomizing devices, for which even most untrained control subjects usually answered with reference to the law of large numbers; to objective problems involving events such as sports, which are readily codable in terms that allow application of the principle and for which many control subjects invoked the principle; to subjective problems about interpersonal relations and other judgments, for which only very few control subjects invoked the principle. Formal training enhanced statistical thinking about equally for these three problem domains.

In a second type of training procedure, *examples training*, subjects were presented with three concrete example problems that illustrated how the law of large numbers could be applied to make inferences about everyday life events. In one study, subjects were presented with example problems and the law of large numbers solutions in one of the three problem domains (probabilistic, objective, and subjective) and then were asked to solve problems in all three domains. Consistent with the formal view, subjects readily extrapolated from the examples: The training effect for problems in the untrained domain was just as great as it was for problems in the trained domain. Training effects were thus domain independent.

Fong et al. (1986) suggested that subjects readily induced abstract rules pertaining to the law of large numbers from the three example problems and thus were capable of applying the principle across domains. There are, however, alternative explanations for the very marked interdomain transfer found in the Fong et al. studies.

One alternative is that the domains used by Fong et al. (1986) were extremely broad. It is possible that the domain independence of training effects was due precisely to the breadth and looseness of the domains. The "domain" of all objectively codable events and the "domain" of all subjective events involving interpersonal relations and other social judgments may be congeries that are too broad to be domains in any meaningful sense. The similarities among problems within a domain may have been so slight that there was no greater ability to apply the rule to problems within the domain than to problems outside it. Research in classification learning (e.g., Fried & Holyoak, 1984; Medin & Schaffer, 1978) has

shown that such learning is facilitated by both the similarity of a particular example to members of the target category and its dissimilarity to members of alternative categories. Thus, to the extent that the three example problems were so variable as to not be well-differentiated from problems in the untrained domain, one would not expect domain specificity effects.

In Experiment 1, we presented subjects with three example problems within more tightly defined domains, sports and ability testing, to explore whether training on more circumscribed content domains would result in reduced transfer effects, as the empirical, antiformalist position would argue. The formalist position, in contrast, would suggest that to the extent that examples training works directly on people's abstract rule system, training on more tightly defined domains would not serve to reduce the transfer to the untrained domain, at least not when testing followed immediately after training.

Our second goal was to address the alternative explanation that training effects were due largely to the use of direct analogies. In the direct analogies account, people map features of a base problem onto those of a target problem, and solution of the target problem is accomplished by direct analogy. As applied to the Fong et al. (1986) results, subjects given examples training may have solved the test problems simply by drawing direct analogies to the example problems. This alternative explanation requires, of course, that subjects remember the example problems at the time they attempt to solve the test problems. In the original Fong et al. studies, testing followed shortly after training, and thus memory for the example problems would have been very great. Because of this, the explanation based on drawing analogies from specific examples cannot be ruled out.

Recent approaches to analogical reasoning have explored the processes by which general principles can be induced from examples (e.g., Dellarosa, 1985; Gentner, 1983; Gick & Holyoak, 1983; Ross & Kennedy, 1990). Gick and Holyoak (1983), for instance, suggested that individuals given two examples abstract a schema based on the similarity and dissimilarity between the elements of the two examples, or base analogs, a process that enhances analogical transfer. Ross and Kennedy (1990) have found that explicit cuing of a prior problem with different content promotes generalizations about how a principle can be applied. They found that such cuing serves to enhance both the access to and the correct use of the principle.

In these studies and others, the details of the prior examples are available and readily accessible in memory because the testing typically occurs very shortly, if not immediately, after training. In the present studies, we wanted to test whether generalization occurred even when details of the examples would be much less accessible and hence more difficult to use in analogical mapping. We did this in Experiment 1 by testing some subjects after a delay of 2 weeks. This delay was intended to degrade memory for the example problems. If training was not maintained at all over the 2-week period, this would be a serious blow to the formalist position, implying that statistical training had no long-lasting effect on improving people's statistical heuristics and that so-called training effects in the immediate testing condition were merely exercises in drawing direct analogies from example problems that were easily accessible. If, on the other hand, there was significant retention of training effects even for the untrained domain, this would suggest that training had its effect through the induction of domain-independent rules. This suggestion would be strengthened by showing that memory for examples was poor after a delay and that individual differences in performance on test problems could not be predicted by individual differences in memory for the details of the example problems.

Experiment 1

Method

Overview of Design and Procedure

In a 2 × 2 factorial design, subjects were trained on the law of large numbers either in the domain of sports or in the domain of ability testing. They were then tested on problems in both domains either immediately after receiving training or after a delay of 2 weeks. In addition, there was a nonfactorial control condition in which subjects received no training before answering the test problems.

Subjects in the training conditions first read the law-of-large-numbers training booklet, which included three example problems. Subjects in the *sports training* condition were given example problems that all dealt with sports in some way. Subjects in the *ability testing training* condition were given example problems that involved sampling a person's mental abilities or intellectual achievements by means of a test or work sample. Each booklet took approximately 15 min to read.

Subjects in the *immediate* condition then answered 10 test problems, 5 in each domain. They were given 45 min to complete the problems. Subjects in the *delay* condition were told that they would return for a second session of the experiment 2 weeks later. They were not told the reason for the second session. The overwhelming majority returned for the second session after exactly 2 weeks, and more than 90% participated within 4 days after that. When delay subjects returned for the second session, they were given the test problems under the same instructions given to subjects in the immediate condition.

Subjects

Subjects were 231 undergraduates at the University of Michigan who participated in partial fulfillment of requirements for their introductory psychology classes. They participated in small groups.

Materials[1]

Training materials. The training booklet began with a one-page introduction to the law of large numbers. The introduction explained that the law of large numbers was a principle of probability that was helpful in understanding and predicting events, especially under conditions of limited information. The introduction then described how the law of large numbers could be used to understand events in one of two domains. In the sports training condition, subjects read the following:

> It is easy to see the application of the principle in the domain of sports. For example, when assessing an athlete's ability, the more

[1] All training and test materials may be obtained from Geoffrey T. Fong.

> games you see him or her play (the larger the sample), the better the idea you get of the athlete's true ability.

In the ability testing training condition, subjects were told the following:

> It is easy to see the application of the principle in the domain of ability testing. For example, when trying to determine whether a person is skillful on a certain task, the more information you have about that person's performance on that task (the larger the sample), the better the idea you get of that person's true ability.

The second part of the training booklet consisted of three example problems. Following each problem was an analysis of it in terms of the law of large numbers. Subjects read each problem and were asked to consider it for a few moments before turning the page to read the law-of-large-numbers answer.

The answers to the example problems were constructed so that subjects could learn how the law of large numbers could be applied to problems in one of the two domains. The example problems and their answers were designed to be structurally identical for both domains. In the sports training condition, the booklet presented three sports example problems. One of these concerned a professional football team that decided to test its policy concerning drafting only players from large colleges. One year, they tried drafting two players from small schools. Both did quite well, and at the end of the year the conclusion was drawn that there was no difference between players from smaller schools and players from bigger schools. The law-of-large-numbers analysis applied to this problem stressed that two players constituted a very small sample of the population. A second problem asked subjects to explain the following phenomenon: After the first 2 weeks of the major league baseball season the leading hitter typically has a batting average as high as .450, yet no batter has ever had an average that high over an entire season. The law-of-large-numbers analysis pointed out that 2 weeks provides a relatively small sample of a batter's ability and that batting averages that are highly discrepant from the average should therefore be more common than they are with a large sample. The third example problem pitted a plausible theory about the deleterious effects of marriage on professional athletes' performance against the conclusions of a controlled study with a large sample that argued against the theory. The law-of-large-numbers analysis emphasized the low probability that the theory could be correct in view of the very large sample of married and unmarried players whose performance was compared, thus illustrating the stability of parameter estimates based on large samples.

In the ability testing training condition, the booklet also presented three example problems. The first described a personnel director who hires flight attendants and who wishes to hire applicants with some knowledge of Spanish. She asks applicants to translate five English words into Spanish. Applicants who do this correctly for four or more words are given priority in hiring. The law-of-large-numbers analysis emphasized that a sample of five words is very small. A second problem described a math teacher who is puzzled about the fact that although he always has two or three students in his calculus class who have averages of 95 or more on the first few weekly tests he gives, no one ever finishes the term with such a high average. The law-of-large-numbers analysis emphasized that extreme scores, both high and low, are to be expected in small samples, because of the high variability associated with such samples. A third problem pitted a plausible theory about the effects of caffeine on intelligence test scores against the conclusions of a well-designed study of the question that used a large number of subjects. The law-of-large-numbers analysis emphasized the stability of large samples.

Testing materials. The testing problem booklet consisted of 10 open-ended problems: 5 sports problems and 5 ability testing problems. The instructions asked subjects to "think carefully about each problem, and then write down answers that are sensible to you. In many of the problems, you may find that the law of large numbers is helpful."

To systematize the kinds of test problems we presented to subjects across the two domains, we constructed the test problems so that each belonged to one of five problem structures (see Fong et al., 1986). Structure 1 problems asked subjects to judge whether a generalization from a small sample was appropriate. Structure 2 problems pitted a small sample against a large sample. Structure 3 problems were regression problems, requiring subjects to explain why an extreme outcome in a small sample was not maintained in a subsequent sample. Structure 4 problems were similar to Structure 2 problems except that the large sample was drawn from a population that was related to, rather than identical to, the target population. Structure 5 problems pitted a large sample against a plausible theory that was offered without any supporting data.

In summary, the 10 test problems followed a 5 × 2 design, with problem structure crossed with problem domain (sports or ability testing). Half of the subjects answered the 5 sports problems first; half answered the 5 ability testing problems first. The order of the problems was randomized within each domain.

Coding System

We developed a coding system similar to that used by Fong et al. (1986) to score the open-ended responses to the 10 test problems. This 3-point coding system was designed to measure the degree to which subjects used statistical concepts such as variability and sample size and whether they correctly invoked statistical principles such as the law of large numbers.

We illustrate the coding system with responses to a test problem used in Experiment 1. In this sports problem, a hockey coach had to decide which of two players, LaBrecque or Stephens, would fill the last spot on his roster. His scouting reports, which were based on five or six games, suggested that Stephens was better than LaBrecque on every dimension. But when the coach watched them for the first time at that afternoon's practice session, LaBrecque played better than Stephens. Subjects were asked which player the coach should choose and why.

Responses to this problem were coded as one of the following three categories:

1 = an entirely deterministic response. Responses in this category included those in which the subject made no use of statistical concepts such as sample size, randomness, or variability. The following was a deterministic response to the hockey problem: "The coach should choose LaBrecque because he played better than Stephens during the practice. Probably the scouts weren't very good judges of ability."

2 = a poor statistical response. Responses in this category included some mention of statistical concepts, but the explanation was incomplete or incorrect. These responses contained one or more of the following characteristics: (a) the subject used both deterministic and statistical reasoning, but was judged by the coder to have preferred the deterministic reasoning; (b) the subject used an incorrect statistical principle, such as Gambler's Fallacy; and (c) the subject mentioned some statistical concept, such as luck or chance, but was not clear about how it was relevant. The following was a poor statistical response: "One practice session is definitely not enough to base a decision on. If only one were allowed, I would choose LaBrecque, because the reports were made by two different people and probably not consistent."

3 = a good statistical response. Responses in this category made correct use of a statistical concept. Some form of the law of large numbers was used, and the sampling elements were correctly identified. In general, the subject was judged to have clearly demonstrated how the law of large numbers could be applied to the problem. The following was coded as a good statistical response: "He should select

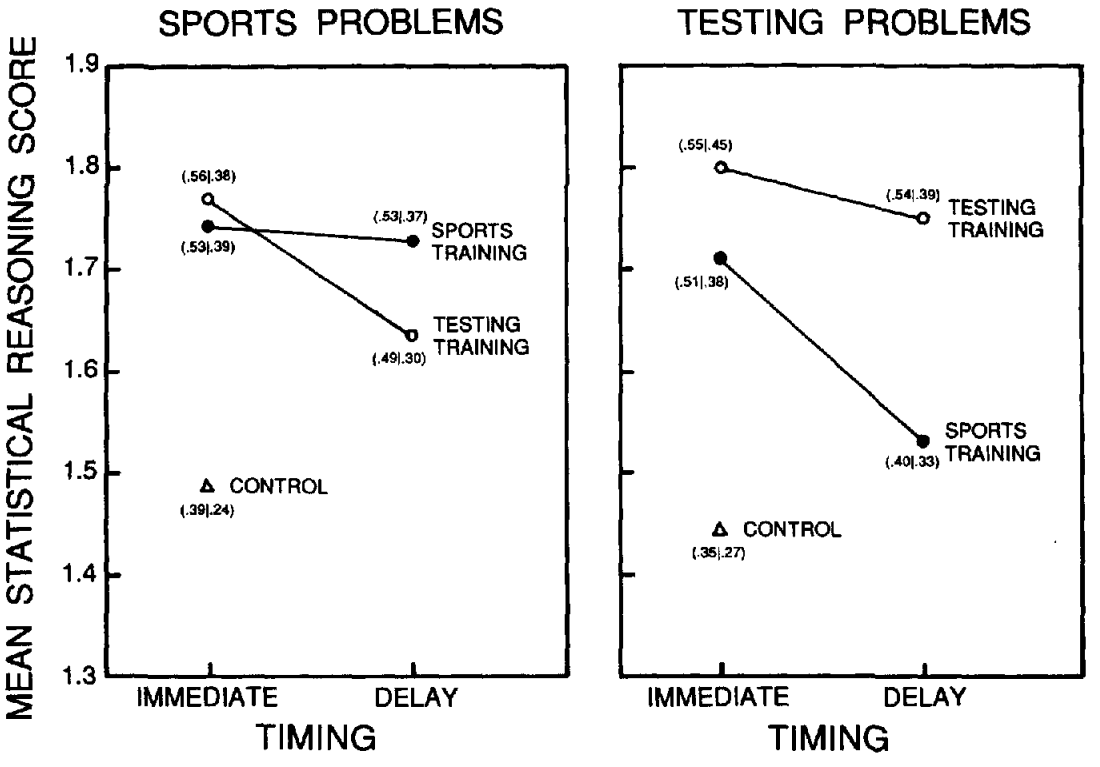

Figure 1. Mean statistical reasoning score as a function of testing domain, training domain, and time of testing (immediate vs. 2-week delay) in Experiment 1. (Within the parentheses is the *frequency* of statistical answers, defined as the proportion of all answers that were statistical in nature [i.e., those answers given a code of *2* or *3*], and the *quality* of statistical answers, defined as the proportion of all statistical answers that demonstrated proper use of the statistical principles [i.e., a code of *3*|*2* or *3*] for each group. Frequency is the proportion to the left of the bar.)

Stephens. The coach had only seen the players once. During that one session, LaBrecque could have had an unusual day, or perhaps Stephens was ill. The scouts saw the players over a larger number of games, and Stephens was the better player."

Because the coding system was designed to identify answers that were statistical in nature, it was possible that a poorly reasoned answer (coded as *2*) would be given a higher score because it mentioned some statistical principle than would a well-reasoned answer that nonetheless failed to incorporate any statistical or probabilistic principle.

Two coders achieved a high degree of reliability: There was exact agreement on 85% of subjects' responses. The statistical reasoning score was computed for each subject, for each of the two problem domains. This was the average score for each subject for the five problems in that domain.

Results

We analyzed the statistical reasoning scores with a modified three-way analysis of variance (ANOVA), using orthogonal contrasts so that the nonfactorial control group could be included. For example, the main effect of the three training conditions was partitioned into the effects of training (average of sports training and ability testing training vs. control) and the effects of training domain (sports training vs. ability testing training).[2]

Figure 1 presents mean statistical reasoning score as a function of testing domain, training domain, and timing of the test problems. Figure 1 also presents, within parentheses, two summary statistics. The first is the *frequency* of statistical reasoning, defined as the proportion of all responses that were coded as statistical, that is, p(code = *2* or *3*). The second is the *quality* of statistical reasoning, defined as the proportion of all statistical responses that were coded as reflecting a good understanding of the principle, that is, p(code = *3*|code = *2* or *3*).

Analyses revealed that there was no effect of the domain of testing; that is, subjects were no more likely to reason statistically for the sports problems than for the ability testing problems, $F(1, 226) < 1$.[3] Apparently, problems in the two domains were approximately equal in the extent to which events could be coded by subjects in terms amenable to intuitive statistical reasoning.

---

[2] There was no main effect for order and no interactions of order with any of the other factors. Thus, order is disregarded in the analyses presented.

[3] The degrees of freedom for all F statistics in Experiment 1 are (1, 226). They are omitted here for the sake of brevity. All p levels reported are based on two-tailed tests.

It may be seen that statistical training had a strong effect on the likelihood that subjects would incorporate statistical concepts in their answers to the test problems ($F = 19.34$, $p < .001$). This replicates the training effects found in Fong et al. (1986). It may also be seen that there was an interaction between training domain and testing domain ($F = 8.90$, $p < .005$). This interaction was dependent on timing. When subjects were tested immediately after training, there was no evidence for domain specificity: The Training Domain × Testing Domain interaction for subjects in the immediate condition was far from statistically significant ($F < 1$). The results thus replicate, for much tighter and more coherent domains, the Fong et al. finding that training effects for the law of large numbers are fully domain-independent when testing is immediate.

It should be noted that the higher statistical reasoning scores for the trained subjects in the immediate condition do not simply reflect a tendency to use the law of large numbers without regard to its correct usage. If that were the case, one would expect that although the frequency of statistical responses would increase, the quality would decrease. That did not occur. In fact, quality, as well as frequency, was higher for the trained subjects.

When subjects were tested 2 weeks after training, domain specificity of training was found, as shown by the significant Training Domain × Testing Domain interaction for subjects in the delay condition ($F = 13.38$, $p < .005$).[4] The three-way interaction (Training Domain × Testing Domain × Timing) was not significant. Thus, the same basic pattern of results—greater loss of training effects over the 2-week delay in the untrained domain—was found for both the sports problems and the ability testing problems.

It is interesting to note that even after a delay of 2 weeks, both frequency and quality of statistical responses were still higher in the trained condition, which is again consistent with the view that subjects induced quite general principles from the examples.

Figure 2 presents (a) the effects of training and timing as a function of whether the testing domain was the same as the training domain or whether it was different and (b) the frequency and quality proportions for each group within parentheses. Figure 2, which simply averages over the specific domains in Figure 1, clearly shows that whereas the effect of training is domain-independent when testing immediately follows training, it is domain-specific after a 2-week delay. As can be seen in Figure 2, performance in the trained domain was unimpaired by a delay of 2 weeks ($F < 1$). In contrast, performance in the untrained domain was significantly lower after the 2-week delay ($F = 12.11$, $p < .001$).

Finally, it should be noted that although there was a significant decline in the retention of training effects in the untrained domain after 2 weeks, there still remained a greater ability to apply the law of large numbers for problems in the untrained domain compared with untrained controls ($F = 4.90$, $p < .05$). This was not a trivial effect. The effect size associated with this contrast, as defined by the standardized mean difference (see Cohen, 1988; Glass, 1976; and Hedges & Olkin, 1985) was 0.44, which is closer to Cohen's (1988) definition of a medium effect size (i.e., 0.50) than it is to a small effect size (i.e., 0.20).[5]

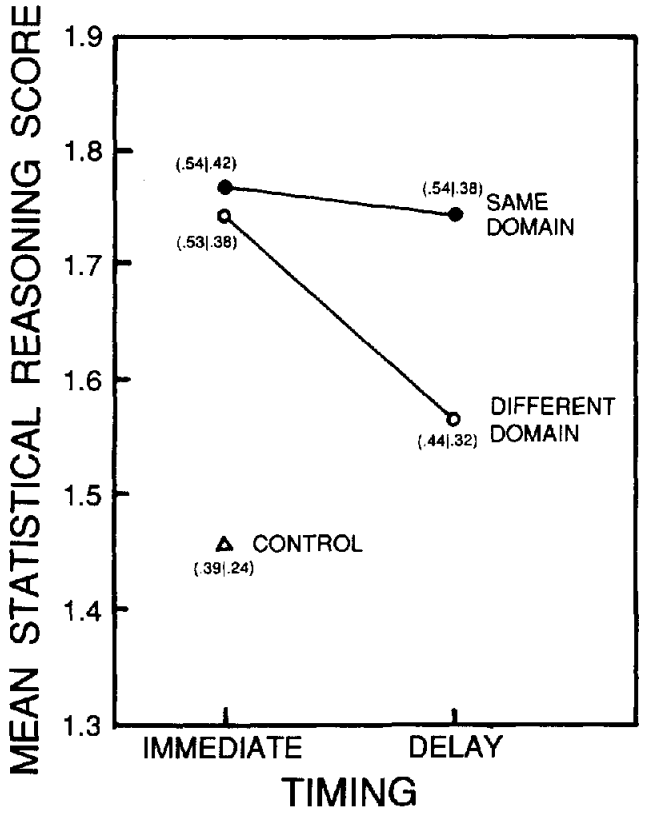

Figure 2. Mean statistical reasoning score as a function of training and testing domain (closed points [●] = performance when testing domain was the same as the training domain; open points [O] = performance when testing domain was different than the training domain) and time of testing (immediate vs. 2-week delay) in Experiment 1. (The frequency and quality of statistical answers are within parentheses.)

Discussion

The results of Experiment 1 make it clear that statistical training effects do not depend on presenting subjects with test problems immediately after training problems. Even when test problems are in a different domain than training prob-

[4] We also conducted analyses on the frequency of statistical answers (see text for the definition of *frequency*). The analyses of frequency were consistent with the analyses reported here.

[5] It should be noted that the difference in statistical reasoning between controls and the condition in which the testing domain was the same as the training domain after the 2-week delay was very strong. The effect size was 1.04, which is well above Cohen's (1988) definition of a large effect size (i.e., 0.80).

lems, there is significant retention of training effects over a 2-week period. We believe that this is due to improvement of a rule system that transcends any given domain in its generality.

How can we account for the very strong retention of training effects in the trained domain? One possibility is that examples training serves not only to enhance the inferential rule system itself, but also to provide subjects with ideas of how to code events within the trained domain in terms of the law of large numbers. These concepts or procedures may be called *coding rules*. For example, subjects given sports examples may learn to view baseball at-bats in terms of samples of a player's ability. Similarly, the intuition that "any team can beat any other team on a given day" can be linked, through examples training, to the formal statistical principle that relates sampling variability to sample size. These coding rules serve to link inferential rules to the content domains in which they are encountered. After a 2-week delay, when memory for the specific examples has decreased, subjects have at their disposal not only the general rules pertaining to the law of large numbers, but also more specific coding rules that aid their performance on the trained domain. These coding rules thus enhance the likelihood that the appropriate inferential rule will be accessed and considered by the individual for application to a given test problem.

This account shares some characteristics with Ross's (e.g., Ross, 1989; Ross & Kennedy, 1990) notion of reminding, in that exposure to earlier examples leads to the induction of general principles that are then invoked when solving new problems. The results of Experiment 1, however, go beyond previous studies in demonstrating that training effects can persevere when subjects are not explicitly cued to the principle and when testing takes place long after training.

Note also that subjects in Experiment 1 were not simply applying the law of large numbers without regard to its correct usage. If that were the case, one would expect that the increase in statistical reasoning scores would be due solely to an increase in the number of poor statistical responses (a code of *2*) and not at all to an increase in the number of good statistical responses (a code of *3*). No such pattern of results was found. Both frequency and quality of statistical responses were higher in trained subjects than in controls.

In addition, there is considerable evidence from Fong et al. (1986, Experiments 1 and 2) that statistical training of the kind used in the present studies did not lead to widespread overuse of the law of large numbers. Fong et al. included in their test package seven false-alarm problems, that is, problems in which application of the law of large numbers was inappropriate for their solution. Fong et al. found that subjects who were given statistical training of the kind used in the present studies were not significantly more likely to invoke the law of large numbers for these false-alarm problems.

Taken together, the results of Experiment 1 suggest that whatever the role of reminding, examples training created a strong and lasting improvement of the rule system itself.

The fact that statistical training effects were maintained with the domain of training over a considerable delay has important implications. It suggests that improvements to reasoning, at least within a fairly broad domain of events, can be very long-standing. This implication is consistent with our view that inferential rules can be abstracted to a high degree from the examples over which they were learned.

Experiments 2 and 3

Our interpretation of the effects of statistical training focuses on the abstraction of statistical inferential rules and coding rules. In this account, statistical training through examples serves both to enhance the abstract rule system corresponding to intuitions about statistical principles and to induce rules about how to code events in the domain represented by the example problems. There is one possible alternative to our theoretical account that suggests that retention of training after a delay results from memory for the example problems themselves and solution of the new problems by analogy. In this view, subjects remember the example problems in sufficient detail that they are able to map the elements of the example problems onto the elements of the test problems they encounter 2 weeks later. Greater retention of training effects in the same domain after a 2-week delay occurs because test problems from the same domain are more likely to remind subjects of the example problems than are test problems from a different domain.

Experiments 2 and 3 were designed to test whether this analogy explanation was plausible by assessing subjects' memory for the details of the example problems after 2 weeks. If memory for the example problems was fairly good, the analogy explanation could not be ruled out. But if memory was poor, then the analogy explanation would be weakened.

The analogy explanation suggests that memory for the example problems should be greater when the test domain is the same as the training domain. In other words, the domain specificity results found for statistical reasoning in Experiment 1 should be manifested in the same pattern for memory of the example problems.

In addition, if, on the one hand, retention of training effects is due to the construction of direct analogies from the example problems, there should be a correlation between statistical reasoning and memory for the details of the example problems. If, on the other hand, it is the memory for the general principle of the law of large numbers that is responsible, there should be a correlation between statistical reasoning and memory for the law of large numbers.

In Experiments 2 and 3, subjects were given examples training in the law of large numbers in either the sports or the ability testing domain. After a delay of 2 weeks, they received either sports or ability testing problems to solve. A third group received no test problems at all. All subjects were then given a questionnaire that assessed their memory for the specific details of the example problems and their memory for the general principle of the law of large numbers. In Experiment 3, there was an additional set of questions that asked subjects who received test problems after the 2-week delay whether the example problems had been helpful in solving the test problems.

Method

Subjects

Subjects in Experiment 2 were 60 members of the University of Michigan subject pool who participated in the experiment in small groups for payment. Subjects in Experiment 3 were 60 introductory psychology students at Northwestern University. There were no differences in the two subject groups; thus, we combined the two groups in the analyses reported here for those measures that were common to both experiments.

Procedure

Subjects were given the same instructions and materials as in Experiment 1. They were given the law-of-large-numbers training materials, which consisted of three example problems, each illustrating the use of the law of large numbers within a certain content domain. These examples were drawn either from the domain of sports or from the domain of ability testing and were identical to those used in Experiment 1. After reading the training materials, subjects were excused and were scheduled to return for the second session 2 weeks later. In other words, the first session was identical in every way to Experiment 1.

During the second session, subjects were assigned to one of three conditions. In two conditions, subjects were given three test problems to solve in which the law of large numbers was relevant. Subjects were given problems in either the sports domain or the ability testing domain. In the third condition, the no test condition, subjects were not given any test problems during the second session.

There were six conditions in all, crossing training domain (sports or ability testing) with testing domain (sports, ability testing, or no test problems).

All subjects then completed a questionnaire assessing their memory for the example problems they had read during the first session. The questionnaire consisted of three pages, each with instructions designed to elicit memory for both the example problems and the law of large numbers.

The instructions on the first page read as follows: "In the space below, we would like you to recall anything you can from the *first* session of this experiment. Provide as much detail as you can."

The instructions on the second page were as follows: "During the first session you read a packet of materials that described some principle. Tell us all you can about the training you received, other than what you already told us on the first page. What was the principle, and what did you read?"

The instructions on the third page were as follows: "In the training materials, you read some example problems. Tell us all you can about the problems, other than what you have already told us on the first two pages. What were the problems about?"

In Experiment 3, the memory questionnaire was lengthened to include a fourth page, which read as follows: "If you have not already done so, please describe how the examples illustrated the use of the principle." In addition, subjects in Experiment 3 were asked whether they had considered the example problems in solving the test problems. For each of the three example problems, they were asked to indicate which of four statements best described their use of the example problem: (a) I didn't think about this example problem in solving today's problems; (b) I thought about this problem, but didn't think it was relevant in solving today's problems; (c) I thought about this problem and tried to use it in solving today's problems; and (d) I definitely used this problem to guide my answers to today's problems.

A coding system was created to assess subjects' memory for the example problems and for the law of large numbers. Codes for each of the three example problems were assigned according to a 4-point scale, on which *1* = the example problem was not remembered at all or nothing was written; *2* = some details about the problem were recalled (e.g., "something about batting averages"), but there was no mention about how the law of large numbers or any other statistical concept was relevant to the problem; *3* = some mention about how the law of large numbers was used in the problem, but the explanation was sketchy, vague, or partly wrong; and *4* = a clear account of how the law of large numbers was used in the problem, that is, how the elements of the problem were relevant to the principle.

This coding for each of the three example problems was performed for each of the three pages of the memory questionnaire in Experiment 2 and for each of the four pages of the memory questionnaire in Experiment 3. In addition, subjects' memory for the law of large numbers or other statistical concepts presented in the training package was coded for each of the three (or four) pages according to a 4-point scale, on which *1* = no description of the law of large numbers, no mention of even its name; *2* = the law of large numbers was mentioned, but without further explanation; *3* = some explanation of the law of large numbers, but sketchy or vague; and *4* = clear explanation of the law of large numbers.

The overall reliability of the coding system was very high: There were exact matches on 88% of the codes assigned by two coders on a subset of the data in Experiment 2 and on 89% of the codes assigned in Experiment 3.

Results and Discussion

Statistical Reasoning

Responses to the open-ended test problems were scored in accordance with the 3-point coding system described in Experiment 1. Table 1 presents the mean statistical reasoning scores as a function of training domain and test domain (omitting, of course, those subjects who received no test problems after the 2-week delay). As can be seen, the domain specificity effects found after a 2-week delay in Experiment 1 were replicated here: The Training Domain × Testing Domain interaction was significant after a 2-week delay, $F(1, 76) = 6.58$, $p < .05$.

Memory for Example Problems

Did memory for the example problems also follow a domain-specific pattern? We analyzed responses to the memory

Table 1
Mean Statistical Reasoning Scores by Training Domain and Test Domain: Experiments 2 and 3

Training domain	Testing domain: Sports	Testing domain: Ability testing
Sports		
M	1.90	1.63
SD	0.49	0.51
Ability testing		
M	1.58	1.83
SD	0.36	0.44

Note. Higher scores indicate greater use of statistical reasoning. Maximum score = 3.0. In each condition, $n = 20$.

questionnaire by using the highest code achieved across the first three pages of the questionnaire for each of the example problems. In this way, subjects would get credit for remembering the details of an example problem whether or not they had done so immediately or after three increasingly detailed prompts (i.e., the questions on each of the three pages of the questionnaire, presented earlier). It should be clear that this procedure was designed to produce the strongest possible memories for the example problems, and the data presented here undoubtedly overestimate subjects' actual spontaneous memory for the details of the example problems.

Having said this, the data show that subjects' memory for the example problems was very poor. Overall, only about one third (35.9%) of subjects recalled any of the three example problems, after all three prompts, with a good understanding of how the problem illustrated the use of the law of large numbers (that is, a code of *4*). Of these, 18.5% recalled one problem, 11.9% recalled two problems, and 5.5% recalled all three.

The mean number of example problems (out of a possible three) that were recalled with a good understanding of the law of large numbers was less than one half of one problem (0.44).

Table 2 presents the mean number of problems recalled with a good understanding of the law of large numbers by training domain and testing domain. A two-way ANOVA demonstrated that there was a significant main effect for testing domain, $F(2, 114) = 3.14$, $p < .05$. Those subjects who were not given test problems after the 2-week delay had a better memory for the example problems that did those who received test problems. Thus, answering test problems after a 2-week delay did not serve to enhance memory for the original example problems, as might be expected if subjects had drawn analogies from the example problems to solve the test problems.[6] In addition, there was no evidence for domain specificity of memory for the example problems, $F(2, 114) < 1$, for the Training Domain × Testing Domain interaction. Thus, when subjects answered test problems from the same domain as the example problems they had read, this did not enhance their memory for the example problems, as would be suggested by the analogical explanation.[7] Finally, there was no difference in memory between sports and ability testing example problems, $F(1, 114) = 1.03$, *ns*.

Table 2
Mean Number of Example Problems Recalled With a Good Understanding of the Law of Large Numbers by Training Domain and Testing Domain: Experiments 2 and 3

	Testing domain		
Training domain	Sports	Ability testing	No training
Sports			
M	0.25	0.25	0.60
SD	0.72	0.55	0.88
Ability testing			
M	0.30	0.45	0.80
SD	0.66	0.76	1.15

Note. Scores are the average number of problems recalled (out of three). In each condition, $n = 20$.

Table 3
Mean Recall Scores for the Law of Large Numbers by Training Domain and Testing Domain: Experiments 2 and 3

	Testing domain		
Training domain	Sports	Ability testing	No training
Sports			
M	3.80	3.35	2.85
SD	0.52	1.14	1.46
Ability testing			
M	3.15	3.75	3.65
SD	1.27	0.64	0.75

Note. Higher scores indicate better memory for the law of large numbers. Maximum memory score = 4.0. In each condition, $n = 20$.

Memory for the Law of Large Numbers

In contrast to the poor memory for the example problems, memory for the law of large numbers was extremely good. More than three quarters (78.5%) of subjects recalled the law of large numbers in sufficient detail and clarity to be given the highest code in our 4-point coding system for law of large numbers memory. If we include subjects who gave at least a fair account of the law of large numbers (code of *3* or *4*), then the percentage goes up to 89.0%. It is clear from these data that subjects remembered the general principle being illustrated in the example problems far better than they remembered the specific example problems themselves.

Table 3 presents the memory for the law of large numbers by training domain and test domain. In contrast to the memory for the example problems, we found a pattern of domain specificity for memory for the law of large numbers. The only significant effect was the Training Domain × Test Domain interaction, $F(2, 114) = 5.37$, $p < .05$. Specifically, memory for the law of large numbers was greatest when subjects received test problems that were from the same domain as they had been trained on 2 weeks previously (the contrast testing this specific effect was significant at the .05 level). This result supports the idea that retention of statistical training effects within the domain of training is partly due to exposure to problems within the same domain; such exposure serves to remind subjects of the general principle of the law

[6] In Experiment 3, a fourth prompt was added that asked subjects to describe how the example problems illustrated the use of the law of large numbers. This prompt elevated the overall average number of example problems recalled with a good understanding of the law of large numbers to only 0.70 (out of a possible 3.0); it did not change the particular pattern of recall by example domain or by test domain.

[7] These data used the most stringent criterion for memory, that is, when subjects were able to remember how the law of large numbers could be used to answer an example problem. If we relax the criterion so that we include subjects whose recall of the example problems with reference to the law of large numbers was only fair to poor (a code of *3* or *4*), the mean number of problems recalled increases to only 0.85 (out of a possible 3.0). Once again, there is no advantage in memory accorded by answering test problems from the same domain as the example problems.

of large numbers better than when they are exposed to problems in some other domain.

Relationship Between Statistical Reasoning and Memory

So far, the results presented demonstrate that the domain specificity pattern for statistical reasoning is mirrored by memory for the law of large numbers, but not by memory for the details of the example problems. This suggests that retention of training effects is due to memory for the inferential rule rather than to memory for the specific examples. However, it still might be the case that those subjects who retained some of the details of the example problems were more likely to use those details to solve the test problems. Thus, a correlation between memory for the example problems and performance on the test problems would support the direct analogy view.

An investigation of the correlations between statistical reasoning and memory, however, discounts this possibility. There was no evidence that memory for the details of the example problems was related to statistical reasoning. The correlation between statistical reasoning and memory for the details of the example problems was only +.06. In contrast, the correlation between statistical reasoning and memory for the law of large numbers was +.31 ($p < .05$).

Reported Use of the Example Problems to Solve the Test Problems

In Experiment 3, subjects were asked whether they had used the example problems in solving the test problems. The most stringent criterion was agreement with the statement, "I definitely used this example problem in solving the test problems." Only 15.6% of subjects agreed with this statement for any of the three example problems. Of these, 9.4% agreed for only one problem, 6.3% agreed for two problems, and none of the 40 subjects who received test problems after the 2-week delay agreed for all three of the example problems. The mean number of example problems used by subjects was only 0.22 (out of a possible 3), and there was no evidence for domain specificity of example problem use, $F(1, 36) < 1$. For instance, subjects receiving sports examples were just as likely to report using the examples to solve ability testing problems as they were to report using them in solving sports problems.

In summary, the data from Experiments 2 and 3 show that (a) subjects had very poor memory for the details of the example problems and how the example problems illustrated the use of the law of large numbers, (b) subjects had very good memory for the general principle of the law of large numbers, (c) the domain specificity pattern for statistical reasoning was echoed in the memory for the law of large numbers but not in the memory for the example problems themselves, and (d) there was a significant correlation between statistical reasoning and memory for the law of large numbers but not between statistical reasoning and memory for the details of the example problems. This pattern of results supports our contention that examples training works by teaching subjects more about the inferential rule system rather than by merely giving them examples to which they can draw analogies when given the test problems. The results also suggest that high retention of training effects is due in part to a greater ability to access the inferential rule system within the trained domain through the use of a rather general set of coding rules linking events in the domain to the inferential rule system.

General Discussion

The results of the experiments reported here are consistent with our view that people make use of abstract inferential rules in reasoning about everyday events and that improvements to these rules can be induced from reasoning about particular examples. In this way, our results are consistent with current views of analogical reasoning that emphasize the induction of general principles from examples (e.g., Dellarosa, 1985; Gentner, 1983; Gick & Holyoak, 1983; Ross & Kennedy, 1990). In contrast, the results are not consistent with the alternative view that people solve such problems exclusively by use of direct analogy.

We found independence of training effects when subjects were tested immediately after training. In this case, reminding certainly played a major role in the effects of training. The example problems were readily accessible in the immediate condition, and analogies were easily constructed. But this same account cannot be readily applied to the results obtained in the delay condition. When subjects were tested 2 weeks after training we found significant retention of training effects even in the domain in which subjects had not been trained. It is highly unlikely that this improvement was produced by direct analogies or remindings of the specific content of the example problems because subjects could recall few details of the examples by that point. In addition, performance was not related to memory for example problems, whereas it was related to the ability to state the abstract rule. Finally, for problems in the trained domain, subjects performed just as well after 2 weeks as they did immediately. If performance was significantly mediated by analogical processes, there should have been some decrement over 2 weeks because of a decline in memory for the example problems.

We suspect that performance on problems within the domain of training was better after 2 weeks than was performance on problems in the untrained domain for two reasons. First, reminding effects (although not for the details of the example problems) were probably at least somewhat responsible: Most subjects would have remembered that they had been told to use the rule for problems in the domain for which they had been trained. However, subjects were more likely not only to use the rule, but also to use it correctly. We believe that they did so by virtue of having learned what we call *coding rules*, which serve to connect the inferential rule system to the domain of training. These rules probably exist at a fairly abstract level, because the test problems differed from the example problems in the kinds of specific content used. Thus, the statistical rules could not have been used to develop answers to the test problems unless the coding rules

were represented at a sufficient level of abstraction to accommodate the many dissimilarities between the example problems and the test problems while still being used by subjects.

In prior studies of inferential rule training, it has not been possible to properly test the alternative explanation that training effects are due to using direct analogies with the actual example problems. By assessing subjects' performance after a delay—when memory for the example problems was greatly reduced, as shown in Experiments 2 and 3—we were able to demonstrate that an explanation based on direct mapping of the example problems onto the test problems is not tenable.

The results are consistent with the assertion that people reason using rules at a high degree of generality and abstraction. A great many scholars today are solidly in the concrete, empirical, domain-specific camp established by Thorndike and Woodworth (1901), arguing that people reason without the aid of abstract inferential rules that are independent of content domain. The present findings—along with those of Fong et al. (1986); Lehman, Lempert, and Nisbett (1988); Nisbett, Fong, Lehman, and Cheng (1987); and Nisbett et al. (1983)—suggest strongly that people do possess abstract rules and that the rules can be improved by methods such as formal instruction (Fong et al., 1986; Fong, Lurigio, & Stalans, 1990; Lehman et al., 1988).

One important implication of our view of inferential rules, and one that is supported in the present studies, is that the effects of inferential rule training need not decay, at least over time periods of days or weeks. When context and content were reinstated, subjects' performances showed the same training increment after a delay of 2 weeks that they showed immediately after training, when the training domain was sufficiently well-defined and codable in terms of the rule system. Indeed, the effect size of training after a 2-week delay was still very large (1.04). Even when the content of the test problems was not the same as that of the training examples, the effect of training was significant after 2 weeks, and the effect size was close to moderate (0.44), as defined by Cohen (1988).

An important task for future research would be to delimit the types of rules and the breadth and nature of the content domains over which such marked retention effects may be found. We note here our strong expectation that such effects will turn out to be more likely for rules that are relatively intuitive and that have counterparts in people's natural intuitive repertoires, such as the law of large numbers (Nisbett et al., 1987). Holyoak, Nisbett, and their colleagues (Cheng & Holyoak, 1985; Cheng, Holyoak, Nisbett, & Oliver, 1986; Holland, Holyoak, Nisbett, & Thagard, 1986) have argued that the laws of formal logic are not included in this category. They have argued that people solve problems, which the logician might solve by applying logical rules, by instead using "pragmatic reasoning schemas." Such schemas are based on the recurrence of certain high regular patterns, for instance, of causal relations, or on the recurrence of contractual relations such as obligations and permissions. These schemas are highly general and abstract, but not so abstract as the purely syntactic rules of formal logic. Consistent with this view, abstract instruction in formal logic was found not to have an effect on the solution of problems involving deductive logic, whereas abstract instruction in pragmatic reasoning schemas did have an effect.

The notion that there is a close relationship between the inferential rule systems that people possess and the extent to which training will succeed helps us to understand why our very optimistic results seem to run counter to the general pessimism of other researchers in the area of transfer of training. Most other transfer studies of problem solving have attempted to teach entirely novel concepts or algorithms, such as the Tower-of-Hanoi problem (e.g., Hayes & Simon, 1977), the missionary-cannibals problem (e.g., Reed, Ernst, & Banerji, 1974), or mathematical rules of probability (e.g., Ross, 1984). When subjects are taught such rules *de novo*, it is likely that it is very difficult for them to induce the appropriate rules at a level of abstraction required for substantial transfer. It is not surprising, therefore, that transfer of training in problem solving leads to a rather pessimistic picture. Although some researchers have demonstrated that teaching completely novel principles can serve to create some degree of generalization (e.g., Ross & Kennedy, 1990), it is an open question whether the effects of training on such novel concepts would be maintained over a long delay, as they were in the present studies.

In the present studies we taught concepts that were already part of subjects' inferential repertoire, and thus the example problems served in part to improve understanding of preexisting inferential rules and in part to provide a set of coding rules that linked the inferential rule system more tightly to events in the trained domain. We believe that such learning should result in potentially better performance in any domain, but especially in those for which coding rules have been formed.

In general, the results of the present studies are more in tune with Plato's abstract view than with the concretism of many twentieth-century views (Nisbett et al., 1987). Rules taught in one domain can transfer to a new domain. And as long as a domain is reasonably tightly defined and coding rules for the domain can be induced, it is possible for the effects of inferential rule training to be retained for a considerable period of time. The results suggest, in fact, that inferential rule training may be the educational gift that keeps on giving.

References

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85,* 1–21.

Braine, M. D. S., Reiser, B. J., & Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 18, pp. 313–371). San Diego, CA: Academic Press.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17,* 391–416.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18,* 293–328.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* San Diego, CA: Academic Press.

D'Andrade, R. (1982, April). *Reason versus logic.* Paper presented at

the Symposium on the Ecology of Cognition: Biological, Cultural, and Historical Perspectives, Greensboro, NC.

Dellarosa, D. (1985). *Abstraction of problem-type schemata through problem comparison* (Report No. 146). Boulder: University of Colorado, Institute of Cognitive Science.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18,* 253–292.

Fong, G. T., Lurigio, A. J., & Stalans, L. J. (1990). Improving probation decisions through statistical training. *Criminal Justice and Behavior, 17,* 370–388.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 234–257.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38.

Gick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier & J. D. Hagman (Eds.), *Transfer of learning: Contemporary research and applications* (pp. 9–46). San Diego, CA: Academic Press.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology, 73,* 407–420.

Hayes, J. R., & Simon, H. A. (1977). Psychological differences among problem isomorphs. In N. J. Castellan, Jr., D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. 2, pp. 21–41). Hillsdale, NJ: Erlbaum.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

Holland, J., Holyoak, K. J., Nisbett, R. E., & Thagard, P. (1986). *Induction: Processes of inference, learning, and discovery.* Cambridge, MA: Bradford Books/MIT Press.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic Books.

James, W. (1890). *Principles of psychology.* New York: Holt.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist, 43,* 431–442.

Manktelow, K. I., & Evans, J. St. B. T. (1979). Facilitation of reasoning by realism: Effect or non-effect? *British Journal of Psychology, 70,* 477–488.

Medin, D. C., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Nickerson, R. S., Perkins, D., & Smith, E. E. (1985). *The teaching of thinking.* Hillsdale, NJ: Erlbaum.

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science, 238,* 625–631.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90,* 339–363.

Reed, S., Ernst, G., & Banerji, R. (1974). The role of analogy in transfer between similar problem states. *Cognitive Psychology, 6,* 436–456.

Reich, S. S., & Ruth, P. (1982). Wason's selection task: Verification, falsification and matching. *British Journal of Psychology, 73,* 395–405.

Resnick, L. B. (1987). *Education and learning to think.* Washington, DC: National Academy Press.

Ross, B. H. (1984). Remindings and their effects in learning a cognitive skill. *Cognitive Psychology, 16,* 371–416.

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 629–639.

Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 42–55.

Thorndike, E. L. (1906). *Principles of teaching.* New York: A. G. Seiler.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review, 8,* 247–261, 384–395, 553–564.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Received June 2, 1989
Revision received August 31, 1990
Accepted September 7, 1990 ■