Journal of Experimental Psychology:
Learning, Memory, and Cognition
1992, Vol. 18, No. 5, 1103-1124

0278-7393/92/$3.00

Role of Analogical Reasoning in the Induction of Problem Categories

Denise Dellarosa Cummins
University of Arizona

In 3 experiments, novices were required to answer questions while reading a series of problems. The questions required them either to analyze individual problem structures (intraproblem processing) or compare problem structures (analogical comparison processing) to derive answers. Ss who engaged in problem comparison processing were found to categorize and describe problems on the basis of problem structures, whereas those who engaged in intraproblem processing, or simply read the problems, categorized and described them on the basis of surface features. Analogical comparisons also facilitated selection and construction of equations relative to intraproblem processing. These results suggest that analogical comparison is an important component in the induction of problem categories.

The purpose of the work reported here was to investigate the role of problem comparison and, specifically, analogical comparison in the induction of problem categories. This work was motivated by two factors. First, it is well-documented that experts and novices represent problems in very different ways and that solution success often depends on producing expert-like problem representations (DeGroot, 1965; Duncker, 1945; Chi, Feltovich, & Glaser, 1981; Hardiman, Dufresne, & Mestre, 1989; Novick, 1988; Schoenfeld & Herrmann, 1982; Silver, 1979, 1981). Second, the problem representations produced by experts and novices appear to reflect differences in the way the two groups organize their knowledge bases. Although both groups appear to represent their problem-solving knowledge in terms of problem classes, or *categories,* expert categories tend to be defined in terms of deep structural features, whereas novice categories tend to be defined in terms of surface features (Adelson, 1981; Chase & Simon, 1973; Chi et al., 1981; Schoenfeld & Herrmann, 1982; Silver, 1979, 1981). Because of this differential organization, experts are more likely than novices to retrieve solution-relevant information from their categories when constructing problem representations.

Category induction, therefore, appears to be intimately involved in expertise development, and a crucial component of this learning appears to be making a transition from relying on surface features to relying on structural features when representing problems and defining problem categories. Despite its importance, this inductive component of expertise development has not received much attention in the literature. Indirect evidence for its involvement, however, can be found in the literature on analogical transfer. In analogical transfer, a learner is taught how to solve a particular problem and is later required to solve another problem that is structurally isomorphic to the original one but has different surface features. If the learner is able to spontaneously recognize the structural similarities between the problems and apply the learned technique successfully, analogical transfer of the learned skill is said to have occurred.

Two important and robust results have been reported in this literature. The first is that when analogical transfer is observed, the learner appears to have induced an abstract schema, or category, that represents the *class* of problems to which the source and target problems belong (Bassok & Holyoak, 1990; Gick & Holyoak, 1983; Holyoak & Thagard, 1989; Novick & Holyoak, 1991; Ross, 1989; Ross & Kennedy, 1990). Surface features appear to be given little weight in the induced problem category representation, and structural features are favored instead. This is a particularly important result because it implies that by *comparing* problems that share deep structures but differ in surface features, novices may come to abstract the crucial problem features that define expert category membership. These problem categories can then, in turn, influence subsequent problem encoding, thereby enabling the novice to represent problems in solution-relevant ways. Analogical reasoning, therefore, may be the key that allows novices to make the transition from relying on surface features to relying on structural features in categorizing and representing problems.

The second robust result reported in this literature, however, is that analogical transfer is notoriously difficult to demonstrate. Typically, novices tend to select prior episodes that share overall surface similarity when constructing solution attempts for transfer problems (Gentner & Landers, 1985; Holyoak & Koh, 1987; Novick, 1988; Ross, 1984, 1987). Very little transfer is observed unless surface feature similarity is maintained between the source and target problems (e.g., Gentner & Landers, 1985), explicit hints are given (e.g., Gick & Holyoak, 1980; Perfetto, Bransford, & Franks, 1983; Ross & Kennedy, 1990), or abstraction is forced by requiring subjects to explicitly cite similarities between the problems (Gick & Holyoak, 1983). Even in the last case,

Experiments 1 and 3 were supported by a grant from the Martin Marietta Corporation, and Experiment 2 was supported by a University of Arizona Small Grant Award to Denise Dellarosa Cummins.

I would like to thank Gary McClelland for statistical advice, Kelly Maurice for programming assistance, and Heather Urry for assistance in data collection. I would also like to thank Brian Ross, Laura Novick, and an anonymous reviewer for helpful comments on a previous version of this article.

Correspondence concerning this article should be addressed to Denise Dellarosa Cummins, Psychology Department, University of Arizona, Tucson, Arizona, 84721. Electronic mail may be sent to dcummins@rvax.ccit.arizona.edu.

transfer does not reliably occur unless learners are allowed to compare at least two problem analogues. Given these bleak results, one may be tempted to conclude that if expertise development depends on analogical reasoning, then it is a wonder that expertise development occurs at all.

From a classification learning viewpoint, this lack of transfer is understandable because successful category induction often depends on the size of the learning set, the variability among exemplars, and the presentation format. Although category induction has been reported after exposure to a single instance (e.g., Elio & Anderson, 1984), the more typical cases show induction as a result of exposure to a sufficiently large number of instances, particularly if there exists a great deal of variability among exemplars (e.g., Homa, Cross, Cornell, Goldman, & Schwartz, 1973; Homa, Sterling, & Trepel, 1981; Homa & Vosburgh, 1976; Posner & Keele, 1968, 1970). This is true in machine learning as well as human learning (Knapp & Anderson, 1984; see also Dietterich & Michalski, 1983, for a review of syntactic methods of bottom-up concept learning). From this perspective, it is understandable why analogical transfer is more likely to be observed when learners have been exposed to more than one exemplar. Multiple exemplars allow the learner ample opportunity to induce a category based on the structural similarities among them. Top-down processes often influence category induction as well by focusing the learner's attention on particular features of the category exemplars at the expense of other features (Carey, 1985; Keil, 1987; Murphy & Medin, 1985). The problem-solving novice's particular reliance on surface features is therefore understandable because naive causal theories of domain-specific phenomena are more likely to be based on knowledge concerning everyday objects rather than abstract theoretical entities. These theories therefore would tend to focus attention on surface feature similarity. Finally, category induction also appears to proceed more quickly and transfer more readily when learning materials are blocked rather than randomly presented (Elio & Anderson, 1981; Homa, 1984). This suggests that requiring learners to compare a reasonably large number of structurally paired exemplars may facilitate category induction, an inference that has not yet been tested in the problem-solving or analogical transfer literature.

There is some indirect evidence to support these conjectures. Although experts often know more about their given domains than novices, the problems used in studies that contrast the two groups typically do not require more domain-specific knowledge than the novices in the studies possess. For example, accomplished junior and senior physics undergraduates know enough physics to solve elementary physics problems. Despite this, physicists and physics graduate students typically outperform them when solving problems (e.g., Chi et al., 1981). When we try to identify the differences between the two groups that could account for the solution performance disparities, it is hard to overlook the fact that novices and experts differ quite dramatically in their degree of experience in solving problems. Physicists have typically encountered and solved many more problems than physics undergraduates. In a similar manner, chess grand masters typically have played many more games (and against more opponents) than chess novices. It is difficult to discount the notion that greater exposure to problem exemplars and problem-solving episodes benefits the learner.

These benefits, however, may simply arise from greater opportunities to analyze individual problem exemplars. Comparing exemplars, analogically or otherwise, may not be necessary to induce solution-appropriate problem categories. Intraproblem analysis (of problem structure, etc.) and greater experience with solving problems may be sufficient to induce useful problem categories. If this is the case, then the results of analogical transfer studies may be misleading. Induction may not be occurring as a result of analogical reasoning, but instead as a simple result of exposure to multiple problem exemplars. To distinguish between these two explanations, direct evidence of the contribution of analogical reasoning to problem category induction is required.

Also of crucial importance is demonstrating induction of categories that are more directly related to the types reported in the expert–novice problem-solving literature. Although this literature tends to use problem tasks chosen from technical domains (e.g., physics, chess, and mathematics), many of the studies investigating analogical transfer use problem tasks that are not characteristic of the expertise literature, such as the Duncker X-ray problem (e.g., Catrambone & Holyoak, 1989; Gick & Holyoak, 1983; Holyoak & Koh, 1987). Those that do use problems that are more closely related to the expert–novice literature (e.g., Bassok & Holyoak, 1989; Novick & Holyoak, 1991; Ross & Kennedy, 1990) typically report indirect measures of category induction (such as written descriptions of problem similarities or anecdotal evidence from verbal protocols) that are unlike those used in the expert–novice literature, where problem sorting and category description are more the norm (e.g., Chi et al., 1981; Schoenfeld & Herrmann, 1982; Silver, 1979, 1981). It was to these ends that the work reported here was conducted.

The materials used were algebra word problems. Word problems were chosen because their problem structures (equations) can be precisely defined, are highly similar to the problem tasks used in the expert–novice literature, and can be solved using a technique (i.e., algebraic manipulation) that was expected to be within the grasp of most college students. The basic methodology involved having students engage in orienting tasks that required them to conduct within-problem analyses or between-problem analogical comparisons. Following this, they were required to categorize old and new (transfer) problems on the basis of problem structure (Experiments 1 and 2) or to select appropriate equations from among alternatives and use them to solve a subset of problems (Experiment 3). This work addressed the following three specific hypotheses: First, it was hypothesized that allowing subjects ample opportunity to compare problem structures analogically, as opposed to analyzing them individually, would facilitate development of problem categories based on structural features. Second, it was hypothesized that induction could occur without the benefit of seeing problem solutions or having subjects solve the problems themselves because comparison of structures was expected to be the core component of category induction. Third, it was hypothesized that categories derived from analogical comparison processes would facilitate later recognition of appropriate solution strat-

egies (equations) because it is problem structure that determines correct solution strategies.

Experiment 1

The purpose of Experiment 1 was to test the hypothesis that analogical comparison processes during problem encoding can lead to category induction. Addressing this issue required a methodology that allowed problem processing to be controlled such that some subjects were encouraged to compare problems, whereas others were hindered or prevented from comparing them. An orienting task methodology was chosen to achieve these ends as nearly as possible. Subjects were required to answer certain types of questions while reading algebra word problems. The questions required them either to compare problems or to search within a problem for the answer. The comparison problems were of two types: those that required the learner to work out analogical correspondences between two problem structures (*analogy*), and those that required the learner to match problem pairs to category descriptions (*schema*). Unlike the intraproblem analysis questions, these questions specifically required, encouraged, and facilitated comparison of problem structures. The intraproblem analysis questions also were of two types: those that drew attention to surface features (*recognition*) and those that drew attention to individual problem structure (*verification*). Note that requiring subjects to answer questions such as these does not completely rule out the possibility of subjects comparing problem structures. There are three reasons, however, to believe that these question tasks make problem comparison less likely than the comparison problem question tasks. First, unlike the comparison tasks, these questions did not explicitly require, encourage, or facilitate comparison of problem structures. Second, as the analogical transfer literature clearly shows, learners rarely notice structural similarities among problems without some kind of help, such as explicit instructions to compare problems. Finally, because the intraproblem processing tasks used here were highly memory-sensitive, optimal performance required that the memory episodes for the problem texts remain distinct. Engaging in problem comparison could compromise memory for superficial and structural detail.

A partial information accretion methodology was used in Experiment 1 such that one group saw only recognition questions, a second saw only verification questions, a third saw verification *and* analogy questions, and a fourth group saw verification, analogy, *and* schema questions. This methodology allowed specific comparisons to be made, which are discussed in detail later. (Recognition questions were not accreted because they imposed too great a memory load for surface details, which overtaxed subjects' memories when included with the other three question tasks.) Following the question orienting task, subjects were required to perform four tasks: (1) Sort the problems on the basis of similarities in surface features, (2) sort the same problems on the basis of problem structure, (3) sort new transfer problems on the basis of problem structure, and (4) describe the problem structures in words.

Because surface features are readily noticed by novices, no differences between the groups were expected on the surface feature sorting task. The remaining three tasks, however, are measures of category induction; moreover, they are exactly the types of category induction measures that are used in the expert–novice literature. Predictions concerning differences in group performance and their implications are as follows:

1. *Verification versus Analogy.* This comparison was the most crucial because it provided a test of the contribution of problem comparison processes to category induction. Given the accretion methodology, both groups performed identical intraproblem structure analysis, but the Analogy group was required to subsequently work out structural correspondences. Because the only difference between these groups is whether the problem structures they analyzed in identical ways were subsequently compared, any differences in group performance must be attributable to the comparison process.

2. *Recognition versus Verification.* This comparison provided information concerning the usefulness of attending to structural information as opposed to surface information for category induction.

3. *Analogy versus Schema.* This comparison provided information concerning the sufficiency of problem comparison processes for category induction. Recall that the only difference between these groups is that the Schema group was given the category descriptions, rather than having to induce them themselves, and therefore were given the opportunity to associate category descriptions with problem instances. Because these subjects were allowed to analyze problem structures locally, compare them, and fit them to category descriptions (with feedback), they should exhibit the maximum level of induction possible under these learning conditions, and hence, the best sorting and describing performance. This group therefore provides a measure of "ceiling performance" on the induction measures. If the Analogy subjects perform equivalently to these subjects, this would mean that they induced categories based on structural features as well as subjects who were actually shown them during training. This result would suggest that analogical comparison processes are sufficient for category induction. It is likely, however, that the Schema subjects would benefit from having actually seen and learned the category descriptions (with feedback) during training. The predicted ordering of the group performance therefore was Recognition < Verification < Analogy ≤ Schema.

Method

Subjects. Forty-eight University of Colorado—Boulder undergraduates participated in Experiment 1 to fulfill an introductory psychology course requirement. The subjects were randomly divided into four groups as described later. The groups were fairly well matched in mathematics experience. The number of subjects in each group having completed at most high school algebra, college algebra, or college calculus, respectively, is as follows: Recognition, *n*s = 5, 2, and 5; Verification, *n*s = 3, 4, and 5; Analogy, *n*s = 5, 3, and 4; and Schema, *n*s = 5, 2, and 5.

Apparatus and materials. Text presentation and collection of responses were controlled by a FORTRAN program run on a PDP 11/03 digital computer. Texts were presented on cathode-ray tube

(CRT) units. Subjects made their responses on response collection boxes located in front of their CRTs.

Twenty-four problems served as experimental stimuli, eight each of three problem structures, catch-up (CU), dilution (DL), and facilitation-interference (FI). A catch-up problem is one in which there is less time than was anticipated to achieve some goal (e.g., earning a certain amount of interest, filling a vat, or traveling a specified distance), and a new rate has to be computed that will allow the goal to be achieved in less time. A dilution problem is one in which a higher rate and lower rate have to be averaged to accomplish the same amount as would have resulted from applying a medium rate for a certain period of time. A facilitation-interference problem is one that involves increasing and decreasing a normal rate of performance by a constant to achieve a constant goal during different lengths of time (e.g., riding the same distance with and against the wind). (See Appendix A for examples of each type of problem structure used in these experiments.)

The eight instances within each problem structure were further subdivided into four topics, which were travel, vat, interest, and work. All problems were nine lines long and ended with a set of three asterisks. Examples of the systems of equations required for solving each of the problem structures are presented in Appendix B. As is apparent, the formulas for solving the problem structures are all variations of the formula, Amount = Rate × Time. However, the variations in problem structure represent nontrivial transformations for novices (cf. Bull, 1982). Moreover, the word problems were sufficiently similar in surface content to render the identification of problem structures a nontrivial task for novices. Compare, for example, these problems with the ones used by Hinsley, Hayes, and Simon (1977). The latter could easily be separated into problem structures by novices simply on the basis of topical word cues that were unique to problems within a given class.

To summarize, problem structure (CU, DL, and FI) was crossed with topic (travel, vat, interest, and work) to produce two instances within each of 12 Structure × Topic cells, for a total of 24 experimental problems. Problems with travel, vat, and interest topics were used during the acquisition stage; the Work topic problems served as transfer problems. In addition, four practice problems were constructed to use when instructing subjects on the orienting and sorting tasks. The algebraic structures of these problems were unrelated to those used in the experimental tasks (i.e., two involved proportions, and two involved two unknowns in two equations), as were the topics (i.e., elections and age). Subjects therefore saw a total of 28 problems during the orienting and sorting tasks, 24 of which were used in the experimental sorting tasks.

The questions used in the orienting task were of four types: recognition, verification, analogy, and schema. Examples of these items are included in Appendix C. All of the questions focused subjects' attention on the same crucial parts of the problems, such as the phrases describing the beginning time and initial rate for filling a vat or earning interest. Recognition questions drew attention to superficial aspects of the texts by requiring the reader to distinguish verbatim repetitions of crucial sentences from truth-preserving paraphrases. For example, consider the following word problem:

> Jill, an aviation technician, is testing a model of an experimental jet plane in a wind tunnel. Flying against the air stream, it takes 10 minutes for the plane to travel the length of the tunnel. Flying with the air stream, it can travel the length in 6 minutes. Jill is amazed at how fast the plane can fly, especially since she knows that she set the wind speed in the tunnel to 15 mph. At what speed can the plane fly with the wind turned off?

An example of a verbatim repetition of a crucial sentence is "Flying against the air stream, it takes 10 minutes for the plane to travel the length of the tunnel." A truth-preserving paraphrase is "Flying against the air stream, *the plane takes 10 minutes* to travel the length of the tunnel" (changes italicized). The recognition task required subjects to respond *same* to verbatim repetitions and *different* to truth-preserving transformations. The verification task required subjects to distinguish between truth-preserving and truth-violating transformations by responding *true* to the former and *false* to the latter. A truth-violating transformation based on the above sentence is "Flying *with* the air stream, it takes 10 minutes for the plane to travel the length of the tunnel" (change italicized).

Analogy questions required learners to work out analogical correspondences, or mappings (A : B :: C : ?), across the same lines of the problems that were used in the verification and recognition tasks. For example, suppose the following problem were blocked with the previous problem during reading:

> For Christmas, Hilda gave each of her two grandchildren an equal amount of money. One child put the money in a certificate. The other put the money in a bond. The rate on the certificate is 2% higher than the regular savings rate. The rate on the bond is 2% lower than the regular savings rate. The certificate earned in 1 year the same amount as the bond did in 3 years. What is the regular savings rate?

An example of an analogy question for the above two problems is the following: "Normal flying speed : Flying with the wind :: Regular interest rate : (a) bond interest rate OR (b) certificate interest rate?" (The correct answer is b.) As this example shows, analogy questions required the reader to compare objects that play the same roles in two problems, thereby allowing them to work out the correspondences between problem structures.

Schema questions allowed learners to associate problem instances with category descriptions by requiring them to select which of a pair of statements accurately described the global problem structure common to two problems. For example, alternative choices for the two problems just cited are as follows: (a) Both problems describe a situation in which a standard rate is increased and decreased by a constant; and (b) both problems involve a catch-up situation, in which there is less time than was anticipated to achieve some goal. The correct answer, (a), is a verbal translation of the equation that is required to solve both problems, that is, $t_1 (r_1 + r_2) = t_2 (r_1 - r_2)$, $t_1 < t_2$, where t = time, and r = rate. Subjects who answered this type of question, therefore, were explicitly shown structural descriptions of the problems and were simply required to learn to associate them with the problems (given feedback).

Procedure. Problems were presented a sentence at a time, and subjects used a button press to request the next sentence. Subjects were tested after each block of two problems. The word "Read" signaled reading blocks, and the word "Ready" signaled testing blocks. Problems were paired such that every possible combination of topics was tested within each problem structure. The same problems were always paired together, but order of problem pair presentation was randomly determined for each subject. The practice problems were always presented first to provide warm-up. Responses were collected and were followed immediately by feedback indicating their correctness.

Subjects were randomly assigned to one of four groups and were tested in groups of four, one to a CRT. The first group (Recognition) received eight recognition-type questions on each block of problems, four for each problem in the blocked pair. The second group (Verification) received eight verification-type questions on each block, four for each problem in the blocked pair. The third group (Analogy) received four verification questions and four analogy questions. The fourth group (Schema) received four verification-type questions, four analogical-type questions, and one schema-type question. The induction measures were taken immediately following the orienting task.

The first sorting task constituted sorting by topic. Subjects were told that the problems they just saw could be categorized into three groups on the basis of similarities in their cover stories or topics.

They were shown an example of how to do this by using the practice problems, grouping Problems A and B together because their stories concerned the age of the characters, and grouping Problems C and D together because their stories concerned elections of candidates to political offices. These were then removed and three more problem cards were placed in front of them. Typed on these cards were experimental problems they also had seen during the orienting tasks. The particular cards chosen differed for each subject, with the constraint that they represented all three topics and structures simultaneously, e.g., Card 1 = travel topic and CU problem structure; Card 2 = vat topic and DL problem structure; and Card 3 = interest topic and FI problem structure. Subjects were told that these three problems represented the three topics by which the remaining problems could be sorted. They were also told that the problems would sort equally, six to a pile, and that they would be given 3 min to sort all of the problems. They were given a few minutes to look the problems over and determine what the three topics were. As soon as all subjects indicated their readiness to begin sorting (usually after less than 1 min), they were allowed to begin. The time left was announced following each minute. When time was up, the experimenter recorded each subject's sorting (codes were printed on the backs of the cards and these were recorded) and took up all of the cards.

Following the topic sorting task, subjects were informed that the problems could be sorted another way, according to similarities in their underlying equation structures, or mathematical principle that describes a fundamental similarity among them. An example based on the four practice problems was shown. Now, Problems A and C were grouped together because their stories both described proportions, and Problems B and D were grouped together because their stories described two unknowns in two equations. They were carefully instructed in how to look for the lines in the problems that described the problem structures (e.g., "6 out of 8 men" in Problem A and "2 out of 3 voters" in Problem C signaled that the two problems both dealt with proportions). Once subjects understood what was meant by problem structures in these problems and how to find them, the practice problems were removed and the same three experimental problems used in the topic sorting task were placed before them. Recall that these problems represented not only three different topics but also three different problem structures. The subjects were informed of this and the fact that, once again, the remaining problems would sort equally into the three piles. The subjects were told that they would be given 5 min to sort the cards and were allowed a few minutes to look the problems over to develop a strategy. During this time, each subject's pile of cards was shuffled. When all subjects indicated their readiness to begin (usually after 1 min), they were allowed to start. The time remaining was announced every minute after the first 2 min had transpired. When time was up, the subjects' sortings were once again recorded, and the sorted cards were taken up.

Subjects were then given the six transfer problems and were told that these, too, could be sorted according to the three problem structures represented by the three example problems. The example problems were displayed again, and subjects were given 5 min to read and sort the new problems into the three piles. Sortings were once again recorded. Following this, subjects were given a protocol sheet on which they were required to indicate the three topics shared by the problems, and to describe the three problem structures by which they sorted the problems on the structure sorting task.

Results and Discussion

In all experiments reported here, significant omnibus F ratios for main effects were followed by planned contrasts, Dunn's post hoc test for paired comparisons when the set of comparisons was constrained, or Tukey's post hoc honestly significant difference (HSD) test for paired comparisons when the set was not constrained. Experiment-wise error for planned contrasts was controlled by using a modified Bonferroni method (Keppel, 1991, pp. 169–170). Significant omnibus F ratios for interactions were followed by simple effects tests (Keppel, 1991). Significant simple effects involving more than one mean were subjected to the paired comparison tests just described.

Orienting task and sorting performance. Subjects performed quite well on the orienting tasks. Average proportions of correct responses and standard errors on each task were as follows: recognition, $M = .62$, $SE = .04$; verification, $M = .74$, $SE = .04$; analogies, $M = .81$, $SE = .03$; and schema questions, $M = .90$, $SE = .04$. The proportion of problems correctly sorted on the topic sorting, structure sorting, and transfer sorting tasks was calculated for each subject. Means for these data are presented in Table 1.

Structure sorting task. The proportions of old and new problems correctly sorted by problem structure were subjected to an analysis of variance (ANOVA) using orienting task (recognition, verification, analogy, or schema) as a between-groups variable, and problem status (new and old) and problem structure (catch-up, dilution, and facilitation-interference) as repeated measures variables.

The analysis returned one significant result, the main effect of orienting task, $F(3, 44) = 12.85$, $MS_e = .21$, $p < .001$. On the basis of the prediction that the ordering of performance would be Recognition < Verification < Analogy ≤ Schema, three planned contrasts were conducted on the group means, using $F(1, 44) = 4.08$, as the rejection value, $p < .05$ (modified Bonferroni method). The results indicated that the Recognition and Verification groups did not differ from each other statistically, ($F < 1$), the Verification group sorted fewer problems correctly than did the Analogy group, $F(1, 44) = 7.60$, $p < .01$, and the Analogy group sorted fewer correctly than did the Schema group, $F(1, 44) = 4.49$, MS_e for all contrasts = .21, $p < .05$. With the exception of the equivalent performance of the two within-problem processing groups, the predictions were supported. The overall pattern of results suggests a clear advantage for problem comparison processing over within-problem processing. The problem comparison groups were well capable of using the knowledge representa-

Table 1
Mean Proportion of Old and New Problems Sorted Correctly on the Basis of Similarities in Topic and Problem Structure in Experiment 1

	Problem sorting criterion		
Processing group	Topic[a]	Structure (old)[a]	Structure (transfer)[b]
Recognition	.98	.40	.53
Verification	.94	.51	.50
Analogy	.98	.67	.76
Schema	.97	.86	.90

Note. Mean proportions are based on 12 subjects.
[a] Each subject's proportion of correct performance is based on 18 problems.
[b] Each subject's proportion of correct performance is based on 6 problems.

tions they constructed during learning to correctly sort both old and new problems; the within-problem processing groups could not.

It is of theoretical importance that orienting task did not interact with problem status ($F < 1$), indicating that the sorting advantage of the problem comparison groups extended to new problems that they had not processed before. Because of its theoretical relevance, it is important to rule out the possibility that this interaction effect was a real effect that was not detected because of small sample size (low power). Partial ω^2 (Keppel, 1991, pp. 223–224) for this effect was less than 1%, indicating that the failure to detect it was not due to low power. This strongly suggests that the comparison groups had formed abstract category descriptions that allowed them to recognize and classify new problem category instances on the basis of problem structure.

Finally, sorting performance was analyzed as a function of orienting task performance. The proportion of all problems (old plus new) correctly sorted by each subject was calculated and entered as the dependent variable in two multiple regression analyses. The predictor variables in the first analysis were (a) proportions of correct answers on the orienting tasks, along with three contrast-coded vectors representing (b) Schema versus Analogy subjects, (c) Analogy versus Verification subjects, and (d) Verification versus Recognition subjects. The second analysis included these variables as well as three vectors representing the interaction of each contrast with the orienting task predictor variable. The former analysis therefore uses an average regression coefficient for orienting task, whereas the latter analysis essentially tests the possibility that the regression coefficients within the groups differed. The first regression accounted for 56% of the variance, $R^2 = .56$, $F(4, 43) = 13.59$, $MS_e = .03$, $p < .001$; adding the interaction terms did not increase R^2 significantly, indicating that the covariate regression coefficients within the groups were statistically equivalent (R^2 change = +.018, $F < 1$). The regression coefficient for orienting task was not significant ($\beta = .17$, $t = 1.23$, $p = .23$), but the coefficients for the three contrasts were (βs = .65, .82, and .49; ts = 4.34, 4.55, and 3.14; ps < .01, for Schema vs. Analogy, Analogy vs. Verification, and Verification vs. Recognition, respectively). Thus, even accounting for individual differences in orienting task performance, the predicted group differences still appeared. Moreover, by reducing variability due to orienting task performance, the predicted difference between the Verification and Recognition groups obtained.

Topic sorting task. The proportion of problems correctly sorted by each subject was subjected to an ANOVA by using orienting task (recognition, verification, analogy, or schema) as a between-groups variable, and problem topic (travel, interest, and vat) as a repeated measures variable. The results indicated that the groups did not differ on the topic sorting task ($F < 1$). Moreover, performance was quite good; the overall mean score was 94%. However, the topics of the problems differed in terms of ease of sorting, $F(2, 88) = 4.83$, $MS_e = .01$, $p < .025$. Tukey's test of pairwise comparisons indicated that the interest problems were significantly easier to sort than the travel problems, whereas the vat problems were not significantly more difficult than either of the other two. (The required difference between mean proportions was .035.) The ordering of difficulty is interesting in that the problems could be viewed as differing in the degree with which they share common entities. For example, the interest problems all contain an investment account, and they were the easiest to sort. All of the vat problems describe a liquid container of some kind, although the nature of the container varied (e.g., bathtub, water bed, etc.), and they followed the interest problems in ease of sorting. The travel problems, however, represented the most heterogeneous set of entities (e.g., plane, jogger, and bicycle), and these were the most difficult problems to sort by topic. This interpretation suggests that subjects notice most readily the similarities among entities in texts.

Written protocol information. It was predicted that the different orienting tasks would affect subjects' internal representations of the problems. To address this issue directly, the category descriptions generated by the subjects were classified according to content. Two independent raters read and sorted the protocols on the basis of the terms that subjects used to described the problem categories. Two categories were decided on in advance. The first (Structural) was to include category descriptions that clearly referred to structural information, such as "There was not enough time to complete a task, and the rate had to be increased." The other category (Surface Feature) was to include protocols that clearly referred to surface or topical aspects of the problems, such as "speed" or "filling." Subjects were to be assigned to a category if at least two of the three descriptions they generated were of the category type (e.g., at least two out of three problem categories described structurally). Fifty percent of the subjects fell indisputably into these two categories. Another 8% were also classified indisputably together and were called the Analytical category. This category consisted of subjects whose descriptions reflected a certain degree of analysis of the problems but an insufficient description of the mathematical structure, such as "different start times, same end times." The remaining 42% of the subjects could not remember the structures, did not answer the question, wrote something unintelligible, or were inconsistent in their descriptions. They were simply classified into a category called Other. The number of subjects in each processing group that fell in each description category is presented in Table 2.

As predicted, the groups differed systematically when describing the three problem structures. Looking down the columns, it is apparent that Schema and Analogy subjects tended to use structural descriptions, whereas Recognition and Verification subjects tended to use surface descriptions. These apparent differences were tested in two ways. First, the relationship present in the entire 4 × 4 table was found to be statistically significant, $G^2(9) = 17.46$, $p < .01$. Second, a portion of Table 2 was extracted and analyzed separately to investigate the relationship between intraproblem and comparison problem processing and reliance on surface and structural sorting strategies. This was done by collapsing the Recognition and Verification groups into a single intraprocessing group, and the Analogy and Schema groups into a single comparison processing group. The number of subjects producing structural and surface feature descriptions in each of

Table 2
Classification of Subjects by Problem Description Generated in Experiment 1

Description	Orienting task				
	Recognition	Verification	Analogy	Schema	Total
Structural	2	1	5	8	16
Analytical	1	1	2	0	4
Surface features	3	3	2	0	8
Other	6	7	3	4	20
Total	12	12	12	12	48

Note. Frequencies are based on 12 subjects in each processing group. Following are examples of descriptions: Structural, "Not enough time, must increase rate"; "Went too fast or too much and must lower rate"; "Equal increases and decreases in rate." Analytical, "Different start times, same end times"; "Involves an average"; "Three rates rather than two." Surface features, "Speed or travel problem"; "Filling or liquids"; "Interest or financial problem." Other, Unclassifiable or no response.

these two groups was then compared. Three intraprocessing subjects produced structural descriptions and 6 produced surface feature descriptions; in contrast, the comparison processing groups produced 13 structural descriptions and 2 surface feature descriptions. This relationship between group type and description type was statistically significant, $G^2(1) = 8.67$, $p < .01$. The groups clearly represented the problem categories to themselves differently. The two groups that were required to compare problems defined the categories in terms of problem structure; the two groups whose attention was focused within problems tended to define them in terms of surface features. The descriptions of the Analogy group are particularly interesting because subjects in this group did not have the benefit of the category descriptions to help them generate structural descriptions as the Schema group did. They, instead, generated these descriptions solely on the basis of their experience with mapping relations across problems during the orienting task.

In contrast with the structure description task, the groups did not differ in the descriptions they generated of the three topics, and all descriptions tended to coincide with surface aspects of the problems (e.g., "travel," "liquids," and "interest").

Summary. Subjects who were required to compare problems while processing them were found to be well capable of sorting old and new problems into categories based on similarities in problem structure and of describing those categories by using structural terms. In contrast, subjects who were required to focus their attention within individual problems while processing them were less capable of sorting problems into structure-based categories. Instead, these subjects relied on surface features while sorting and describing problems.

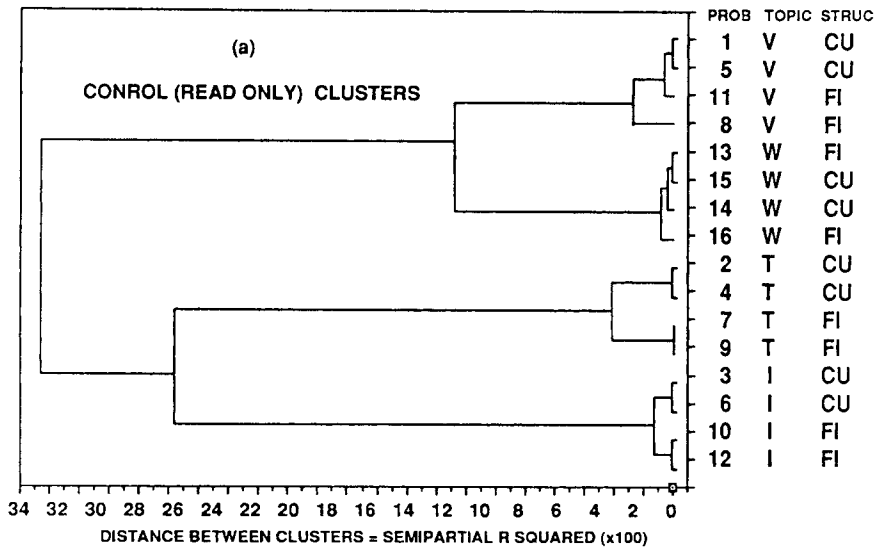
Experiment 2

The purpose of this experiment was two-fold. The first was to rule out an "elaborative processing" explanation of the results of Experiment 1, that is, the possibility that the superior performance of the Analogy and Schema groups was due simply to benefits derived from answering multiple types of questions and hence laying down multiple memory traces. The second concerned the importance of analogical mapping during learning relative to schema matching. The benefits of exposure to category descriptions alone could not be ascertained because the Schema group answered analogy questions as well. Results of other studies suggest that learners find it easier and more informative to construct solutions from concrete examples than to base them on abstract, schematic instructional materials, at least early in the learning process. For example, when first learning to write LISP functions, it is often easier to model one's code on example functions than it is to work from function definition templates (Pirolli & Anderson, 1984, 1985). Programmers who have reached an intermediate level of expertise, however, typically find it more convenient and efficient to use the templates when writing functions. The learning process seems to proceed from analogical mapping to schema mapping over the course of expertise development. This would suggest that the high level of performance demonstrated by the Schema subjects may have been largely due to their experience with answering analogy questions (which allowed them to compare concrete examples) rather than their experience with mapping problems onto category descriptions.

Addressing these two issues simply required having groups of subjects perform orienting tasks that consisted of a single type of question: verification, analogy, or schema. As mentioned earlier, Gick and Holyoak (1983, 1987) found that requiring subjects to explicitly describe similarities between problems facilitated category induction. A particularly important comparison, therefore, was the one between the Schema group and the Analogy group. The Schema subjects were given the opportunity to explicitly state the global similarities between the problems, just as subjects were in Gick and Holyoak's investigations. The Analogy subjects, however, were given the opportunity to explicitly map structural correspondences.

In addition to the three orienting task groups, a fourth group was used that simply read the problems before sorting them. This group constituted a control group, and their performance was expected to provide a baseline measure of category induction when subjects are left to their own devices. It was predicted that baseline would fall somewhere between the Verification and Analogy group performance levels for the following reasons: The results of Experiment 1 clearly suggest that attending primarily to intraproblem information (as required by the verification task) hinders category formation relative to comparing crucial structural correspondences between problems (as required by the analogy task). When left to their own devices, reasoners are free to compare problem structures. As the analogical transfer literature indicates, however, they are often not very good at this unless they are given explicit hints as to where to look. As a result, the control group was expected to lag behind the Analogy group (who were given explicit information as to where to look for useful comparisons) but exceed the Verification group (who were distracted from making such comparisons).

Finally, rather than sorting by topic, all subjects in Experiment 2 were required to perform a free sort of the problems,



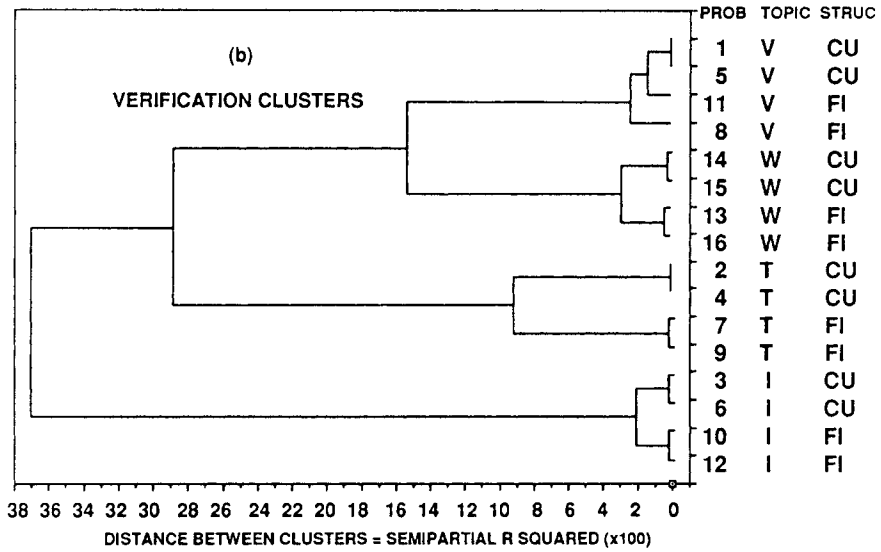


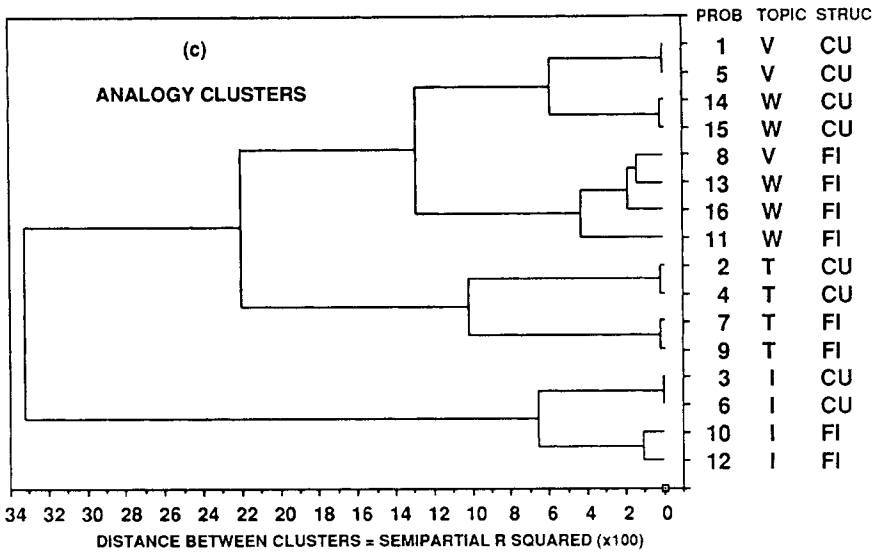
Figure 1a–1b. How novices sorted algebra word problems after answering verification questions or only reading the problems. (Topics: V = vat; W = work; T = travel; I = interest. Structures: CU = catch-up; FI = facilitation-interference.)

putting all problems together that they believed could be solved by using the same specific equation or procedure.

To summarize, subjects either simply read a series of problems (read-only) or engaged in one of three orienting tasks while reading (i.e., verification, analogy, or schema). They then sorted the problems into as many categories as they believed necessary to capture the similarities in solution procedures. Finally, they sorted old and new problems into a fixed number of categories for which examples were shown. The expected ordering of performance levels was Verification $<$ Control $<$ Schema $\leq$ Analogy.

Method

Subjects. Seventy-two University of Arizona—Tucson undergraduates participated in Experiment 2 to fulfill an introductory psychology course requirement. The subjects were randomly divided into four groups as in Experiment 1. Scholastic Aptitude Test (SAT) quantitative scores were obtained from subjects as a measure of mathematics ability. (Scores were not available for subjects who were transfer students because the registrar's office does not record them.) An ANOVA was performed on the available scores by using Group as a between-subjects variable. The average SAT quantitative scores for the Verification (n = 15), Analogy (n = 17), Schema (n = 15),



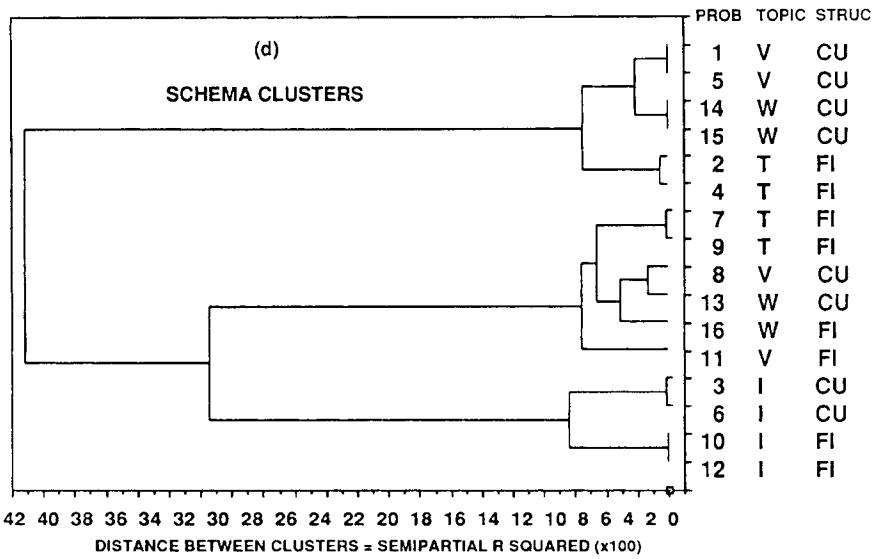


Figure 1c–1d. How novices sorted algebra word problems after answering analogy or schema questions. (Topics: V = vat; W = work; T = travel; I = interest. Structures: CU = catch-up; FI = facilitation-interference.)

and control ($n = 14$) groups were 540, 592, 510, and 590, respectively. These means did not differ from each other, $F(3, 57) = 2.56$, $MS_e = 12{,}170$, $p > .05$, indicating that the groups were equivalent in mathematics ability.

Apparatus and materials. The same stimuli used in Experiment 1 were used here, except that the dilution problems were eliminated. This meant that the number of problems used was reduced from 24 experimental and 4 practice problems (Experiment 1) to 16 experimental and 4 practice problems, and the number of categories was reduced from three (Experiment 1) to two. This was done to reduce the time requirements of the study. The orienting task was presented as a pencil-and-paper task. The presentation appeared as it had on the computer in Experiment 1, with each screenful of information appearing on a separate page in a booklet. Subjects were told not to look back or forward through their booklets in order to keep the presentation as close as possible to the computerized presentation in Experiment 1. For each problem pair in the booklets, the Verification subjects saw four verification questions, the Analogy subjects saw four analogy questions, and the Schema group saw one schema question. No subject saw multiple types of questions. The practice problems always appeared as the first two pages in the booklet, and the remaining pages were randomized for each subject.

Procedure. Subjects were tested in groups of four to six. They signed a consent form that requested their SAT scores and gave the experimenter permission to verify their scores with the admissions office. The orienting task booklets were then distributed, and the subjects were given 20 min to complete them. Feedback was given concerning the correctness of their answers after they completed the entire booklet.

Following this, they were given cards on which the experimental problems were typed, along with forms for recording their sorting decisions. The forms were standard sheets of white paper that were divided into sections labeled *Category 1, Category 2,* and so on, up to *Category 16.* They were instructed to sort the problems on the basis of similarities in solution procedures, putting all problems together that could be solved in the same way, that is, by using the same equation or other solution procedure. They were told to feel free to sort the problems into as many categories as they needed and to write a description of each category in the spaces provided. Under each description, they were to record the members of the category by listing the problems' identifier codes found on the back of the cards. They were given 30 min to complete this task. Following this, they were required to perform the same structure sorting task that was used in Experiment 1, except that only two categories were used instead of three. The procedures were the same otherwise. These forms and cards were removed and the transfer task was performed, using the same instructions and procedures as in Experiment 1 except that two rather than three categories were used.

Results and Discussion

Free-sort task. A hierarchical cluster analysis was performed on the sortings that subjects produced by using the cluster analysis procedure in the SAS statistical package. Several sorting algorithms were used, including Ward's minimum variance procedure, the centroid hierarchical procedure, McQuitty's similarity analysis, and the complete linkage method. Because they all produced essentially the same cluster results, only the results of Ward's method is reported here. Ward's method was chosen because, similar to regression procedures, its measure of error (i.e., variability among or distance between observations) is a sum of squared deviations, which is readily converted to proportion explained variance. In Ward's method, each object, initially, is itself considered a cluster. Cluster pairs that are closest to each other (i.e., have the smallest sum of squared deviations) are collapsed into a single cluster. In this way, average squared distance between objects within a cluster is minimized relative to the average squared distances between clusters. This collapsing of clusters continues in stepwise fashion until all of the objects ultimately are clustered into a single group. In terms of variance accounted for, the procedure begins with all variance accounted for (i.e., each object assigned to its own cluster, hence variability between objects completely accounted for) and ends with zero variance accounted for (i.e., all objects assigned to a single cluster, hence variability between objects completely unaccounted for). Thus, each time the clusters are collapsed into a smaller number of clusters, the proportion of explained variance decreases. When choosing the best solution, it is customary to choose the cluster solution stage at which further collapsing of clusters results in an intolerable loss of explained variance. (See Dunn-Rankin, 1983, p. 139, for further details.)

To begin the analysis, subjects' sortings were converted to percentage overlap scores for each possible pair of problems. (See Dunn-Rankin, 1983, pp. 41–48, for a complete description of how this is done.) An overlap score indicates how often two problems were sorted into the same category. This frequency is divided by the number of subjects who performed the sorting to produce a proportion. This proportion is a measure of how similar two problems are, because it is assumed that similar problems will be categorized together more often than dissimilar problems. By subtracting this proportion from one, the similarity measure is transformed into a dissimilarity measure. These dissimilarities were used as inputs to the SAS cluster analysis procedure.[1]

Figures 1a through 1d depict the cluster solutions at each stage of the analysis. The length of the lines in Figures 1a through 1d depict the distance between the clusters in terms of semipartial squared correlation units; they can also be thought of as representing the change in variance accounted for at each stage of the cluster analysis—the longer the line, the greater the change in explained variance. The changes in explained variance as a function of number of clusters is more explicitly depicted in Figures 2a through 2d.

The contents of the clusters indicate which problems were sorted together most frequently and hence were judged as being most similar. Figures 1a and 1b show the results of the clusters produced by the control and Verification groups, respectively. By scanning the figures from right to left, the clusters formed during each stage of the cluster analysis can be seen. Looking at the far right-hand side of these figures, it is apparent that both of these groups sorted the problems into four clusters, and that these clusters corresponded to the four problem topics: Vat, Work, Travel, and Interest. Figures 2a and 2b show clearly that proportion explained variance asymptotes at the four-cluster stage, indicating that this division provides the optimal solution for these groups. This four-cluster solution accounted for 91% and 81% of the variance in the control and Verification group sortings, respectively. By working inward from the left side to the right side of Figures 2a and 2b, major divisions in the analysis can be seen. For the control group, the first division occurs between Interest and Travel problems on the one hand, and Work and Vat problems on the other. The second major division splits these two clusters into four clusters—Interest, Travel, Work, and Vat. For the Verification group, the first division is between Interest problems on the one hand and all other problems. The second division splits Travel from the Work and Vat problems, and the third division splits Work and Vat problems. As will be seen, all subjects seemed to assign special status to Travel and Interest problems, presumably because

[1] The SAS procedure assumes coordinate data or distance data based on coordinates, and the one minus percentage overlap measure of distance used here was considered appropriate for this procedure for the following reasons: (a) Unlike multidimensional scaling techniques, cluster procedures have been shown to be robust to violations of assumptions that typically underlie distance calculations, that is, symmetry, triangle inequality, and the distance between x and y being zero only if $x = y$ (Shepard, 1980); (b) Dunn-Rankin (1983, p. 139, paragraph 4) states that, according to a personal communication from Ward, percentage overlap data is used more often than distance data for job classification and task analyses.

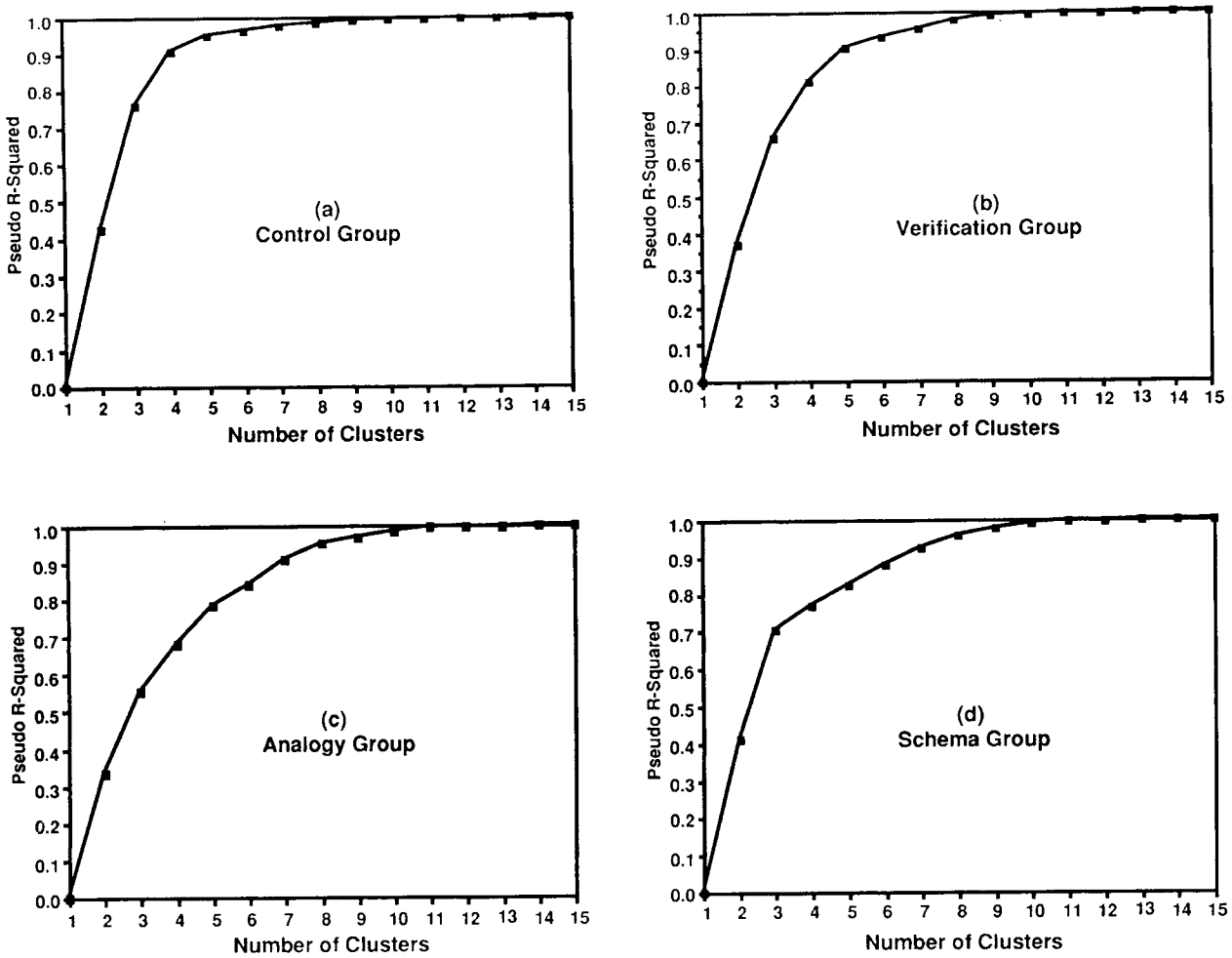


Figure 2. Changes in proportion of explained variance as a function of number of clusters in cluster analysis solutions for the Verification, Analogy, Schema, and control groups.

they remember that there are special equations for dealing with these types of problems (even if they cannot remember what those equations are), whereas the same cannot be said for Work or Vat problems. Interest problems also seem the easiest to separate from the rest, as the major divisions show. As mentioned earlier, the Interest problems all made reference to an investment account, hence subjects' propensity to cluster them may reflect the degree to which their surface forms contain common entities. Note also that these subjects were not entirely insensitive to problem structure. The smallest subdivisions (far right-hand side of Figures 2a and 2b) show that, in many cases, problems were paired on the basis of similarities in both topic and structure (e.g., Travel CU problems form one subdivision, and Travel FI problems form another). Despite this, structural information played virtually no role at higher stages of the clustering analysis and is entirely absent at the final (and optimal) solution stage of four clusters.

Skipping ahead to the clusters produced by analysis of the Schema group's sortings (Figures 1d and 2d), a very different picture emerges. Here, acute sensitivity to problem structure can be readily seen. Figure 2d shows an abrupt rise in proportion explained variance from one to three clusters, followed by a gradual increase until eight clusters, where asymptote is reached. The three-cluster solution accounts for 71% of the variance, and the eight-cluster solution accounts for 96%. The reason for this striking difference in the Schema group's clustering solution becomes apparent when Figure 1d is compared with Figures 1a and 1b. Looking at the far right-hand side of Figure 1d, the apparent three-cluster solution shows that two of the clusters formed by the Schema group are based solely on problem structure and only one is based on topic. This is in striking contrast with Figures 1a and 1b, where the clusters are clearly defined by topic alone. The top cluster on the right side of the Figure 1d contains all of the CU problems except Interest ones. The middle cluster contains all of the FI problems except Interest ones. The third cluster contains the Interest problems. The special status of Interest problems can also be ascertained from the major divisions. Working inward from the left-hand side of Figure 1d, the first division can be seen to separate all non-Interest CU problems from all other problems. The second division separates Interest problems from the others. The third division

separates the CU Interest problems from the FI Interest problems, indicating sensitivity to problem structure even within this topic cluster. The divisions also make clear that the Schema group was not insensitive to surface features when sorting the problems, and this accounts for the gradual increase in explained variance as one moves from a three-cluster to an eight-cluster solution. For example, consider the CU problem cluster in the upper right-hand side of Figure 1d. Note that the subsequent divisions of this cluster are made on the basis of topic lines. Vat CU problems form a separate subdivision within the cluster, as do Work CU problems and Travel CU problems. The gradual increase in explained variance between three and eight clusters (Figure 2d) is consistent with the claim that, although Schema subjects were acutely sensitive to problem structure, their sortings were influenced by surface features as well. Recall that the 16 problems consisted of two instances each of four topics crossed with two problem structures. If similarities in both problem structure and topic were used during sorting, this would produce eight categories containing two problem exemplars each. It is clear that the bulk (71%) of the variability in the Schema group's sortings is accounted for by sensitivity to structural features, with the remaining 25% (to asymptote) indicating sensitivity to differences in surface features.

Turning now to the results depicted in Figures 1c and 2c, the Analogy group's sortings can be seen to fall somewhere between those of the control and Verification groups and those of the Schema group. Like Figure 2d, Figure 2c shows large changes in proportion explained variance from one to three clusters (55%), and smaller changes from three to eight clusters, at which point asymptote is clearly reached (96%). Notice, however, that the rise from one to three clusters is not as great for the Analogy group as for the Schema group (55% vs. 71%) and that the subsequent increase to asymptote is far more gradual and smooth for the Analogy group than for the Schema group. The reason for these differences in explained variance become apparent when comparing Figures 1c and 1d. Looking at the far right-hand side of Figure 1c, four clusters are immediately apparent, two of which are defined by structure and two by topic. Scanning down the right-hand side of Figure 1c, it is apparent that the first cluster contains all Vat and Work catch-up problems, and the second contains all vat and work facilitation-interference problems. The remaining two clusters contain travel and interest problems. Working from the left side of the Figure 1c toward the right, the first division can be seen to separate Interest problems from other problems, the second separates Travel problems, and the third distinguishes between CU and FI problems among those that remain. A six-cluster solution distinguishes between CU problems and FI problems within the Interest and Travel clusters (at the third major division), and an eight-cluster solution distinguishes among the Vat and Work problems within the FI problems. Thus, although both the Schema group and the Analogy group were sensitive to both structural and surface features, comparison of Figures 2c and 2d shows a difference in the degree of sensitivity. The Schema group appears to have weighted structural features a bit more heavily than surface features, whereas the Analogy group appears to have weighted the two types of information more evenly. The most important result of these free sort cluster analyses, however, is the striking sensitivity to structural features exhibited by the problem comparison groups that is nearly absent in the sortings produced by the control and Verification groups.

Another important thing to notice about the cluster trees is how the Work problems were sorted by the different groups. These problems were not used during the orienting or reading tasks and hence are transfer problems. As is apparent, the control and Verification groups clustered these problems together into a single category. The Analogy and Schema groups, however, integrated them into higher order clusters based on similarities between their equation structures and the equation structures of the other problems. Thus, when given new problems, these groups were capable of responding to the problems' structural features when making classification decisions.

Orienting and structure-sorting tasks. Once again, subjects performed well on the orienting tasks. Mean proportions of correct responses and standard errors were as follows: schema, $M = .84$, $SE = .03$; analogy, $M = .89$, $SE = .02$; and verification, $M = .97$, $SE = .01$. The proportion of old and new problems correctly sorted by each subject into structure-defined categories (when category examples were provided) were analyzed by ANOVA. Group (Schema, Analogy, Verification, or control) served as a between-groups variable and problem structure (catch-up and facilitation-interference) and problem status (old and new) served as repeated measures variables. The means from this analysis are presented in Table 3. As predicted, the main effect of group was significant, $F(3, 68) = 4.49$, $MS_e = .12$, $p < .01$. As in Experiment 1, it is important to note that problem status did not interact with group, indicating that the group differences reported later were reflected in both old-problem and new-problem sortings, $F(3, 68) = 1.03$, $MS_e = .03$, $p > .05$. Partial ω^2 for this effect was less than .001, indicating that the failure to detect this interaction was not due to a power problem.

Five planned contrasts were conducted on the group main effect, using $F(1, 68) \geq 4.91$, $p \leq .03$, as the significance value (modified Bonferroni method). The first three compared the processing groups with the control group. The results of these contrasts indicated that the Analogy group surpassed the

Table 3
Mean Proportion of Old and New Problems Sorted Correctly on the Basis of Similarities in Problem Structure in Experiment 2

Processing group	Problems	
	Old[a]	Transfer[b]
Read-only	.50	.54
Verification	.51	.51
Analogy	.68	.69
Schema	.67	.61

Note. Mean proportions are based on 18 subjects.
[a] Each subject's proportion of correct performance is based on 12 problems.
[b] Each subject's proportion of correct performance is based on 4 problems.

control group when sorting problems on the basis of problem structure, $F(1, 68) = 8.42$, $MS_e = .12$, $p < .01$; the Schema group differed marginally from the control group, $F(1, 68) = 4.11$, $MS_e = .12$, $.08 > p > .05$; and the Verification and control groups performed equivalently ($F < 1$). Again, given the small sample size used in the Experiment 2, the question arises as to whether the nonsignificant results of the latter comparisons could be due to a power problem. Keppel (1991, pp. 129–130) recommends an estimate of the strength (or magnitude) of a comparison based on the ratio of the sum of squares of the comparison to the sum of squares of the comparison plus the sum of squares of the error term ($R^2 = SS_{comp}/SS_{comp} + SS_{error}$). This measure comes close to providing an estimate of the effect size that might be found in a two-group experiment involving the comparison. The R^2 values for the above comparisons were .11, .06, and .001, for the Analogy–Control, Schema–Control, and Verification–Control comparisons, respectively. Thus, it seems reasonable to conclude that the marginal results of the Schema–Control comparison resulted from low experiment power, whereas the failure to detect a difference for the Verification–Control comparison reflects an effect size that is negligible. These results suggest that processing tasks that require learners to compare problem structures facilitate identification and abstraction of such structures. Relative to the control and Verification groups, the Schema and Analogy groups were better at categorizing and describing problems on the basis of equation structure.

Comparing the Analogy group to the Verification and Schema groups underscores even more dramatically the importance of guided structure comparison. Analogy subjects surpassed the Verification subjects on this sorting task and equalled the Schema subjects, $Fs(1, 68) = 9.10$ and .76, $R^2 = .12$ and .01, respectively. Thus, even though these subjects were not given cohesive, higher order descriptions of the structures during training, they were well capable of constructing them by making specific structural comparisons between pairs of problems. Recall that the Verification subjects were shown the same pairs of problems and were required to answer questions about the same lines in the problems. The only difference between the Verification and Analogy groups was whether their orienting task questions focused their attention within the problems or required them to compare the problems. Problem comparison clearly facilitated recognition and representation of problem structures.

It is also interesting to note that the equivalent performance of the Analogy and Schema subjects in Experiment 2 resulted from a drop in performance among the Schema subjects relative to Experiment 1. Overall (old and new combined), the performance of the Analogy (and Verification) subjects remained steady across the two experiments, with Analogy subjects sorting an average of 69% of the problems correctly in Experiment 1 and 68% correctly in Experiment 2. (The corresponding figures for the Verification groups were 51% in both experiments.) Average performance for Schema subjects dropped across the two experiments from 87% in Experiment 1 to 65% in Experiment 2. This pattern suggests that the Schema subjects in Experiment 1 benefited from working out the structural correspondences highlighted by the analogy questions they answered as well. In contrast, the Analogy subjects in Experiment 1 did not seem to derive any additional benefits from answering verification questions, presumably because being required to compare two problem structures is redundant with answering questions that require analyzing them separately. Consistent with the results of other studies on skill acquisition, therefore, it appears that optimal performance obtains in learning conditions that provide the learner some exposure to abstract definitions *and* opportunities to analyze concrete examples. These results suggest, furthermore, that it is the opportunity to *compare* problem structures, as opposed to simply analyzing them individually, that most benefits performance.

Also significant were the effects of problem status and the Problem Status × Problem Structure interaction, $Fs(1, 68) = 4.80$ and 6.38, $MS_e s = .24$ and .10, $ps < .05$, respectively. Simple effects indicated that old CU and FI problems were sorted with equivalent ease ($F < 1$) but that new FI problems were significantly easier to sort than new CU problems, $F(1, 68) = 5.92$, $MS_e = .30$, $p < .025$.

Finally, as in Experiment 1, an analysis of sorting performance was conducted by using orienting task performance as a covariate. Two multiple regression models were used, one that included orienting task as a predictor along with contrast-coded vectors for the Schema–Analogy comparison and the Analogy–Verification comparison, and one that included the two-way interactions of orienting task with these contrast vectors. (The control group's performance could not be included because they performed no orienting task.) The first regression model accounted for 21% of the variance, $F(3, 50) = 4.36$, $MS_e = .03$, $p < .01$, and the second model did not account for appreciably more, (R^2 change = +.024 .1, $F(2, 48) = .91$, $MS_e = .02$, $p > .05$. Closer inspection of the results of the simpler model indicated that the regression coefficient for the Analogy–Verification contrast was significant ($\beta = .59$, $t = 3.56$, $p < .01$); and the coefficient for the Schema–Analogy contrast was not ($\beta = .25$, $t = 1.69$, $p > .10$). The coefficient for orienting task was marginal ($\beta = .27$, $t = 1.86$, $.05 < p < .07$). These results therefore corroborate the results of the raw data, indicating that the contrast of interest (Analogy–Verification) was significant even when individual differences on the orienting task are taken into account.

To summarize, as in Experiment 1, subjects in Experiment 2 who compared problems during processing tended to sort them on the basis of underlying problem structure, whereas those who focused attention on individual problems (or were left to their own devices) tended to sort them on the basis of surface similarity. This was true when subjects were allowed to sort the problems into any number of categories as they wished, and when they were instructed to sort them into a fixed number of categories for which examples were given.

Structured Sort Analysis: Protocols

As in Experiment 1, subjects were classified on the basis of the problem category descriptions they produced. The same four categories were used here (i.e., structural, analytical, surface feature, and other). The frequencies are presented in Table 4. The relationship was again tested in two ways. First,

Table 4
Classification of Subjects by Problem Description Generated in Experiment 2

Description	Orienting task				Total
	Read-only	Verification	Analogy	Schema	
Structural	1	0	5	9	15
Analytical	5	5	4	1	15
Surface features	8	5	5	2	20
Other	4	8	4	6	22
Total	18	18	18	18	60

Note. Frequencies are based on 18 subjects in each processing group. Following are examples of descriptions: Structural, "Not enough time, must increase rate"; "Went too fast or too much and must lower rate"; "Equal increases and decreases in rate." Analytical, "Different start times, same end times"; "Involves an average"; "Three rates rather than two." Surface features, "Speed or travel problem"; "Filling or liquids"; "Interest or financial problem." Other, Unclassifiable or no response.

the relationship between processing group and description type present in the entire 4 × 4 table was found to be significant, $G^2(9) = 8.88$, $p < .01$. Second, the Read-Only and Verification groups were collapsed into a single group and compared with the Analogy and Schema subjects on structural and surface feature descriptions. Three Read-Only and Verification subjects produced structural descriptions and 13 produced surface descriptions; 14 Analogy and Schema subjects produced structural descriptions and 7 produced surface descriptions. This relationship was also significant, $G^2(1) = 6.93$, $p < .01$. The subjects who engaged in comparison processing showed a clear tendency to produce structural descriptions when sorting the problems, whereas the intraprocessing and control subjects tended to produced surface descriptions.

Experiment 3

The purpose of Experiment 3 was to investigate the influence of problem comparison processes on the selection and execution of solution strategies. The same three orienting tasks in Experiment 2 were used here. In addition, there was a novice–control and a group of "experts" who simply read the problems before the experimental tasks. Subjects' tasks in Experiment 3 were to match old and transfer problems to equations, set up the equations for a subset of the problems, and manipulate the equations to derive solutions. The important question was to what extent the categories induced through problem comparison (particularly analogical mapping comparison) facilitate the recognition and construction of symbolic problem representations. In mathematics and science, it is of crucial importance that concrete problem situations be translated into symbolic representations that are manipulated syntactically. For example, when solving a physics problem, knowledge of physical principles is used to translate verbal or pictorial problem representations into equations (i.e., symbolic representations). These equations are then manipulated syntactically (e.g., algebra, calculus, arithmetic) to derive a symbolic solution. The symbolic solution is then translated back into a verbal or pictorial form, or used to guide some action. If category induction indeed underlies expertise development, one would expect the categories induced through comparison processes to facilitate the first part of this process, namely, the selection and construction of a symbolic representation. The remaining parts are beyond the scope of these processes and this article because individual differences in subjects' algebraic skills could not be controlled.

The experimental procedures were straightforward. Following the orienting tasks or reading period, subjects were shown three solution procedures that laid out the algebraic structure of the three problem structures (catch-up, dilution, and facilitation-interference). They were required to match old and transfer problems to these solution procedures, set up the equations for a subset of them by assigning values from the problems to the variables in the equations, and manipulate the equations to derive a solution. The predicted order of performance levels on the equation selection and set up tasks were as follows: Verification < Novice-Control < Schema < Analogy = Experts.

Method

Subjects. Sixty subjects were recruited from campus newspaper advertisements for participation in the experiment. Experts consisted of 8 mathematics graduate students, 1 physics graduate student, 1 molecular biology graduate student, and 2 seniors in computer science. The majors of the novices were many and varied; however, there were no preponderances of engineering, mathematics, or nonscience majors in any particular group. All subjects were paid $7.50 for their time.

Materials and apparatus. The same problems and apparatus were used in Experiment 3 as in Experiment 1. The three solution procedures corresponding to the three problem structures are included in the Appendix B. The three solution procedures were typed on separate sheets of paper. Solution procedures for the two types of practice problems were also constructed.

Procedure. All subjects were told before reading the problems that they would be required to match some algebra word problems to a set of solution procedures and to solve a subset of the problems by using the procedures. The novices were randomly assigned to four groups: a Verification group, an Analogy group, a Schema group, and a control group. The first three groups went through the same orienting task procedure as in Experiment 1. The control group and the experts simply read the problems on cards before sorting them.

Control novice subjects and experts were given the stack of cards with the problems typed on them and given as much time as they needed to read them. Most required 15–20 min to read all 18 problems and 4 practice problems. When subjects had finished reading the problems (or performing the orienting task), they were shown how to match the practice problems to their appropriate solution procedures. Specifically, this meant pointing out which numbers should be assigned to which variables and then working through the algebra in some detail to obtain the answers.

Following this, the three solution procedures for the rest of the problems were distributed. These were explained in detail but no attempt was made to explain how to match the problems to the solution procedures. For the catch-up problem solution procedure (referred to as Solution 1), subjects were told that for this type of problem, there was a goal rate that was associated with a certain period of time and a higher rate that was associated with a shorter period of time. Applying the rates for their respective periods of time produced the same goal amount. For the dilution problems (referred to as Solution 2), there was a goal rate that was associated with a certain period of time, a higher rate that was associated with a shorter

period of time, and a lower rate that was associated with a longer period of time. By combining the results of the higher rate at the shorter time and the lower rate at the longer time, one could arrive at the same amount as the goal rate at the goal length of time. For the facilitation-interference solution procedure (referred to as Solution 3), there was a standard rate and another rate. Combining the two rates allowed the same amount to be accomplished in a shorter period of time than would be the case if the second rate were subtracted from the first rate.

When all subjects indicated that they were ready to begin, they were allowed to begin matching the problems to the appropriate solution procedure by laying the cards on the bottom of the solution procedure sheets. They were given 10 min to complete the sorting. When they were finished, their sortings were recorded. The problems were stacked in one pile face down on the table. Then they were given the six transfer problems and were asked to read and sort these new problems. They were given 5 min to complete the transfer sorting. Their sortings were recorded, and the transfer problems were added to the stack.

Following the sorting task, the subjects were then given sheets of paper on each of which was written a problem number and the number of the solution procedure they had assigned that problem. They were instructed to solve the problems using that solution procedure. They were told that if they changed their minds and wanted to use another procedure, to do so but to record the number of the solution procedure they were now using. It was stressed that they should (a) write down the values they associated with each variable, (b) form equations as shown, and (c) manipulate the equations to obtain their answer, boxing their final answer.

Each subject was required to solve six problems, one old problem, and one transfer problem from each problem structure. Problems were assigned to subjects in advance such that all problems were attempted by an equal number of subjects in all groups; problem order was random for each subject. Subjects were given 30 min to solve the problems and were notified every 10 min of the time left.

Results and Discussion

Orienting task. Once again, orienting task performance was quite good. Mean proportion of correct performance and standard errors were as follows: $M = .72$, $SE = .04$, on the verification task; $M = .85$, $SE = .02$, on the analogies; and $M = .89$, $SE = .04$, on the schema question task.

Matching task. The proportion of problems correctly matched to solution procedures on the first sorting task and the transfer task was calculated for each subject. The means for these data are presented in Table 5. An ANOVA was conducted with the following variables: group (verification, analogy, schema, expert, or novice-control), problem status (old and new), and problem structure (CU, DL, and FI), with repeated measures on the last two variables. The main effect of group was significant, $F(4, 55) = 9.04$, $MS_e = .18$, $p < .001$, as was the main effect problem structure, $F(2, 55) = 15.95$, $MS_e = .03$, $p < .001$. These variables interacted with each other, $F(8, 110) = 2.10$, $MS_e = .03$, $p < .05$. Simple effects indicated that the interaction resulted from the fact that experts matched all three problem structures equivalently ($F < 1$), whereas all of the novice groups (with the possible exception of the Analogy group) found CU problems significantly easier to match to their equations than DL or FI problems, $Fs(2, 110) = 10.78$, 2.38, 7.40, and 3.52; $MS_e = .03$; $ps < .001$, .10, .001, and .05, for the Verification, Analogy, Schema, and control groups, respectively. Because the ordering of the group means within each problem structure was the same, the comparisons of interest were conducted on the overall group means. The relevant statistics for the Group × Problem Status interaction were as follows: $F(4, 55) = 1.96$, $MS_e = .10$, $p = .12$, partial $\omega^2 = .007$.

Table 5
Mean Proportion of Old and New Problems Correctly Matched to Equations in Experiment 3

Group	Problems	
	Old[a]	Transfer[b]
Read-only	.52	.65
Verification	.46	.47
Analogy	.65	.71
Schema	.78	.64
Expert	.84	.90

Note. Mean proportions are based on 12 subjects.
[a] Each subject's proportion of correct performance is based on 18 problems.
[b] Each subject's proportion of correct performance is based on 6 problems.

To address all questions of interest, nine comparisons were conducted. The critical difference between means required for Dunn's test at the .05 level was .206. The first four comparisons were concerned with comparing novice performance with expert performance. The observed differences between the Expert group mean and the means of the Verification, control, Analogy, and Schema groups were .403, .285, .192, and .162, respectively. Thus, the Verification and control groups differed significantly from the Expert group when matching problems to equations, whereas the Analogy and Schema groups did not. Caution should be used when interpreting these results, however, because of the small sample sizes used in Experiment 3. The strength of these comparisons (as described in Experiment 1) were as follows: $R^2 = .37$, .23, .12, and .09, for the Verification, control, Analogy, and Schema comparisons, respectively. Thus, the Analogy–Expert and Schema–Expert comparisons appear to reflect nontrivial effect sizes that could not be detected because of low power. They are much smaller, however, than the effect sizes of the other two comparisons. A more conservative interpretation of these results, therefore, is that the schema and analogy orienting tasks *reduced* the differences between novices and experts on the sorting task.

The next three comparisons were concerned with comparing the orienting task groups to the novice-control group. The observed differences between the control group and the Verification, Analogy, and Schema groups were .118, .092, and .123, respectively. Because the required difference was .206, none of these comparisons was significant. Again, caution should be used when interpreting these results. The effect sizes for these comparisons were as follows: $R^2 = .05$, .03, and .05, suggesting that larger sample sizes may have allowed these effects to be detected.

The final two comparisons concerned differences between the orienting task groups. The observed difference between the Verification and Analogy group means was .211, and

between the Schema and Analogy group means was .030. Thus, the Analogy group matched reliably more problems to their proper equations than did the Verification group, whereas the Analogy and Schema groups did not differ significantly (R^2 = .14 and .003, respectively).

As in the other two experiments, the sorting results of the Schema, Analogy, and Verification groups were analyzed in multiple regression analyses that allowed orienting task performance to serve as a covariate. (The Verification group contained two outliers that artificially inflated R^2; the covariate and sorting group means were substituted for these outliers in the analyses that follow.) The first analysis, which contained only the covariate and the group contrasts, accounted for 52% of the variance, $F(3, 32) = 11.77$, MS_e = .02, $p < .01$, and adding the covariate–contrast interaction vectors did not produce a significant increase in R^2, R^2 change = .07, $F(2, 30) = 2.51$, $MS_e = .02$, $p > .05$. The regression coefficient for the Analogy–Verification group contrast was significant, as was the coefficient for orienting task (β = .58, and .32; t = 3.48 and 2.22; ps < .002 and .03, respectively). The coefficient for the Schema–Analogy contrast was marginal (β = .30, t = 1.92, .05 < p < .06). Once again, the results of the covariance analysis corroborate the results of the analysis of raw scores.

Equations. As mentioned earlier, novices were not chosen for their facility with manipulating equations, making comparison of final solutions between the novice and expert groups uninformative. The more informative measure of understanding is how well the novices performed relative to experts in terms of setting up equations. If they set up an equation properly but could not derive a correct solution, this would mean that they understood the problem's mathematical structure and could represent it adequately but could not produce a correct answer because their algebraic skills were rusty or lacking.

Subjects' equation protocols were scored as follows: For each variable assigned a correct value, one point was given. The maximum possible score for the CU and FI problems was 3, because there were four variables in each problem, three whose values were given. The maximum score for the DL problems was 5, because there were six variables, five whose values were given. Each subject was required to set up equations for two problems of each type, one old problem and one transfer problem. Their scores were converted to proportions. The mean proportion equation scores for the five groups are presented in Table 6. An ANOVA was computed on these data by using group as a between-subjects variable, and problem structure and problem status (new or old) as within-subject variables.

The analysis returned three significant results, the main effects of group, problem structure, and problem status, $F(4, 55) = 3.59$, $MS_e = .24$, $p < .025$; $F(2, 110) = 4.27$, MS_e = .13, $p < .025$; and $F(1, 55) = 6.95$, $MS_e = .11$, $p < .025$, respectively. The relevant statistics for the Group × Problem Status interaction were as follows: $F < 1$, partial $\omega^2 < .01$.

Once again, each novice group was compared with the expert group by using Dunn's test, the required difference between mean proportions being .24. The observed differences between means were as follows: Expert–Verification = .24, Expert–Analogy = .17, Expert–Schema = .24, and Expert–Control = .26. These results indicate that the expert group performed significantly better than each of the novice groups *except* the Analogy group. As in the matching performance analysis, however, caution should be used in interpreting these results because of the small sample sizes used in Experiment 3. R^2 values for these comparisons were .14, .07, .14, and .16, respectively. Thus, the Analogy–Expert contrast may constitute a real effect that was not detected because of low power. The more conservative interpretation of these effects is that making analogical comparisons reduced the difference between novice and expert performance.

Table 6
Proportion of Correct Equations Constructed From Problems in Experiment 3

Group	Problems	
	Old[a]	Transfer[b]
Read-only	.70	.66
Verification	.76	.64
Analogy	.81	.73
Schema	.78	.61
Expert	.96	.93

[a] Proportions are based on 12 subjects × 18 problems = 216.
[b] Proportions are based on 12 subjects × 6 problems = 72.

The significant main effect of problem status suggests that the problems used as transfer problems in these experiments were significantly more difficult to process than the experimental problems. Because this variable did not interact with group, this suggests that even the experts and novice-control subjects (who received these problems as just another stack to read) found them troublesome. To verify this null effect, a difference score was computed for each subject (proportion old minus proportion new), and contrast weights of −3, +2, +2, +2, and −3 were assigned to the means from each group. This contrast was not significant, although the direction of the means was suggestive of no real differences for the expert and novice-control groups, and an attenuated difference for the Analogy group, $F(1, 55) = 2.56$, $MS_e = .08$, .15 > p > .05; Verification = .13, Analogy = .09, Schema = .17, expert = .01, and novice-control = .02.

The main effect of problem structure indicated that they were not all equally easy to solve. Significantly lower scores were obtained on the FI problems than on the CU problems, with the DL problems falling between and not differing from either. The interaction of group and problem structure was not significant ($F < 1$); even the expert group found the FI problem type to be tricky.

Solutions and errors. For the sake of completeness, the proportion of problems that were correctly solved by each subject was calculated by ANOVA, using group (novice-control, verification, analogy, schema, or expert) and problem status (old and new) as variables, with repeated measures on the latter variable. Mean proportions are presented in Table 7.

The main effect of group was significant, $F(4, 55) = 4.81$, $MS_e = .16$, $p < .01$. Using Dunn's test (required difference between means = .30), the expert group was found to have

Table 7
Mean Proportion of Correct Solutions in Experiment 3

Processing group	Mean proportion
Novice-control (Read-only)	.43
Verification	.36
Analogy	.47
Schema	.42
Expert	.78

Note. Each subject's proportion of correct solutions is based on 6 problems. Mean proportions are based on 12 subjects.

produced significantly more correct answers than each of the novice groups. This result is understandable given that the expert group was chosen from fields in which algebraic manipulation of equations is an everyday affair, whereas most of the novices in this study rarely used algebra in their respective study disciplines. A different picture emerged, however, when the types of errors committed were analyzed. Errors were scored as either conceptual or mechanical, where a conceptual error meant using the wrong equation or assigning wrong values to the variables, and a mechanical error meant making an arithmetic or algebraic manipulation error. The means for conceptual errors are presented in Table 8. The groups were found to differ from each other, $F(4, 55) = 2.65$, $MS_e = 1.68$, $p < .05$. Using Dunn's test (required difference = 1.39), the novice-control group was found to have committed significantly more conceptual errors than the Expert group ($R^2 = .12$), and the difference between the Verification and expert groups was marginal (V − E = 1.34, $R^2 = .10$). Analogy and Schema subjects, however, did not differ from the expert subjects. Unlike the Verification–Expert comparison, this was probably not due to low power. R^2 for the Schema–Expert and Analogy–Expert comparisons were .01 and .02, respectively. The overall pattern of results of the equation setup, solution, and conceptual error analyses suggest that requiring subjects to explicitly compare problem structures improved their understanding of these structures, relative to novices who were not told to compare them, and attenuated the differences between them and expert problem solvers.

Summary of sorting performance across the three experiments. Table 9 presents the mean proportion correct structure sorting performance for old and new problems combined in each experiment, along with their respective standard errors. The stability of the manipulations is apparent in Table 9. When left to their own devices, novices appear to sort a little over 50% of the problems correctly, whether the sorting requires matching problems to each other or to equations. The same is true for novices who are required to verify structural information within problems. Requiring novices to compare problem structures analogically, on the other hand, raises performance to somewhere around 70%. Adding a requirement to compare structures with abstract descriptions (schema questions) appears to improve performance to over 80% (Experiment 1). Requiring them to compare structures with abstract descriptions without the benefit of analogical comparisons also appears to improve performance relative to leaving them to their own devices, but the measures for this manipulation appear somewhat less stable (Experiments 2 and 3). Taken together, however, the stability of these measures indicate that the experiments replicate each other well and that a reasonable degree of confidence can be taken in the experimental manipulations.

Table 8
Mean Number of Conceptual Errors Committed During Solution Attempts in Experiment 3

Processing group	Mean errors
Novice-control (Read-only)	1.83
Verification	1.75
Analogy	0.83
Schema	1.00
Expert	0.42

Note. Maximum possible errors = 6. Means are based on 12 subjects.

General Discussion

The results of these experiments clearly demonstrate the importance of analogical comparison processes in the induction of problem categories. In all three studies, subjects who were required to compare problems consistently sorted and described them in terms of problem structures, whereas those who were required to analyze them individually (or were left to their own devices) tended to sort and describe them in terms of surface feature similarity. The benefits of problem comparison were also seen on the equation selection, setup, and solution tasks. Here, subjects who engaged in problem comparisons committed no more conceptual errors than did experts when attempting to solve problems. The difference between expert and novice performance on the equation selection and setup tasks was also attenuated by requiring novices to compare problem structures. These results indicate that the knowledge acquired from mapping structures between problems greatly enhances the learner's ability to recognize symbolic representations of problem structures and to map values from natural language texts onto their appropriate mathematical expressions.

More important, these results strongly suggest that this inductive component is crucial for developing expertise in a domain. Subjects who engaged in analogical comparison seem to have developed a better understanding of the problems, as demonstrated by the clarity of their problem descriptions and their ability to construct coherent mathematical representations of the problem structures. Other researchers have also reported the importance of analogical mapping in learning (Mayer, 1975, 1976, 1983; Mayer & Bromage, 1983; Mayer & Greeno, 1972). For example, when learning to write LISP recursive functions, students typically model their coding attempts on example solutions, modifying the examples as needed to meet the demands of the new problem statement. In a similar vein, Lewis and Anderson (1985) reported that subjects who were allowed to learn a recursive procedure by discovery (bottom-up induction from examples) were better able to recognize errors.

Finally, these results suggest that expertise development may rely on the same inductive processes that underlie concept formation in general. One would expect, therefore, that

Table 9
Summary of Combined Sorting Task Performance (Old + Transfer) From Experiments 1, 2, and 3

	Proportion of total problems correctly sorted by structure					
	Experiment 1[a]		Experiment 2[b]		Experiment 3[a]	
Processing group	*M%*	*SE*	*M%*	*SE*	*M%*	*SE*
Read-only	—	—	.51	.03	.55	.06
Verification	.51	.05	.51	.02	.46	.05
Analogy	.69	.05	.68	.05	.70	.06
Schema	.87	.05	.65	.05	.73	.04

[a] Each subject's proportion of correct performance is based on 24 problems. Mean proportions are based on 12 subjects. [b] Each subject's proportion of correct performance is based on 18 problems. Mean proportions are based on 18 subjects.

expertise development would be sensitive to the same variables that influence the acquisition of other classes of information. There is some evidence that this is true. The heart of the bottom-up component in classification learning is the detection of covariances among features in category exemplars. Features that occur frequently across examples are included in the concept definition, and those that vary are viewed as irrelevant. This means that nondefining features that happen to occur in many instances may become part of one's category description (e.g., *flying* and the concept *bird*). The same is true for problem domains. Zhu and Simon (1987) required students to learn factorization procedures from worked-out examples. They reported that bottom-up inductive learning can take place with little knowledge of a domain, but it requires exposure to numerous examples and may lead to erroneous learning if irrelevant features are consistently present in the examples. Bassok and Holyoak (1989, 1990) found more transfer from algebra to physics problems than vice versa. Their results suggested that this differential transfer was due to the degree and nature of content-embedding in the two types of problems. Physics principles are typically content-specific (e.g., accelerating masses), and their embedded concepts are complex rather than unitary. Algebraic principles, on the other hand, are typically not tied to any specific content and their embedded concepts are typically unitary. Variability among examples has also been found to slow initial problem category learning, just as it does when learning to classify other materials (Gick & Holyoak, 1987).

Category induction is also sensitive to top-down influences, such as the learner's prior knowledge of the domain, sensitivity to causal structure, and naive theories of domain-based phenomena (Carey, 1985; Keil, 1987; Murphy & Medin, 1985). These top-down influences often focus the learner's attention on particular features of the category exemplars at the expense of other features. This has the ultimate effect of constraining the set of features over which an inductive generalization is to be made. The accuracy of the resulting generalization depends on the accuracy of the knowledge-derived feature bias. The differences between expert and novice problem classifications fit nicely into this framework. Experts draw their constraints from general knowledge of the field, whereas novices may draw their constraints from naive theories concerning the relationship between surface and structural features (Medin & Ortony, 1989). As a result, experts tend to constrain their generalizations to structural features at the expense of surface ones, whereas unassisted novices tend to constrain their generalizations to surface features (Chi et al., 1981; Hardiman et al., 1989; Novick, 1988; Ross & Kennedy, 1990). One could conjecture that problem-solving experience benefits the novice by providing feedback (solution success and failure) that can be used to modify the differential weightings on the two types of information, leading ultimately to a shift of focus to structural features.

In support of this view of expertise development is evidence that novices often are not blind to structural information nor are experts blind to surface features. Hardiman et al. (1989) presented physics experts and novices with a series of problem triads consisting of a model problem and two comparison problems. Their task was to select which of the two comparison problems could be solved in the same way as the model problem. The comparisons matched the model problems on the basis of surface structure, problem structure, both, or neither. The results indicated that experts relied more on problem structure than did novices, although they, too, were negatively influenced by surface feature overlap. Moreover, novices differed in their degree of sensitivity to problem structure, and these differences correlated with later problem-solving performance. Novick (1988) also reported that mathematics experts showed more positive transfer between structurally analogous problems than did novices, and less negative transfer between problems that had only surface features in common. Even in these studies, however, the influence of both types of information is apparent. For example, in Experiment 3, 46% of the expert group exhibited negative transfer (and, hence, sensitivity to surface feature overlap), and 29% of the novice group exhibited positive transfer (and, hence, sensitivity to structural feature overlap). These results suggest that the difference between the two groups may be a matter of differential *weighting* of surface and structural information in the knowledge base rather than the presence or absence of the two types of information.

References

Adelson, B. (1981). Problem solving and the development of abstract categories in programming languages. *Memory & Cognition, 9,* 422–433.

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 153–166.

Bassok, M., & Holyoak, K. J. (1990). Transfer of domain specific problem-solving procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 522–533.

Bull, K. (1982). Expert and novice differences in solving algebra word problems. Unpublished doctoral dissertation, University of Colorado at Boulder.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1147–1156.

Chase, W., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152.

DeGroot, A. D. (1965). *Thought and choice in chess.* The Hague: Mouton.

Dietterich, T. G., & Michalski, R. S. (1983). A comparative review of selected methods for learning from examples. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (pp. 41–82). Palo Alto, CA: Morgan Kaufmann.

Duncker, K. (1945). On problem solving. *Psychological Monographs, 58,* Whole No. 270.

Dunn-Rankin, P. (1983). *Scaling methods.* Hillsdale, NJ: Erlbaum.

Elio, R., & Anderson, J. R. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 7,* 397–417.

Elio, R., & Anderson, J. R. (1984). The effects of information order and learning mode on schema abstraction. *Memory & Cognition, 12,* 20–30.

Gentner, D., & Landers, R. (1985). Analogical reminding: A good match is hard to find. *Proceedings of the International Conference on Systems, Man, & Cybernetics* (pp. 607–613). Tucson, AZ.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306–355.

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 and J. D. Hagman (Eds.), *Transfer of training: Contemporary research and applications* (pp. 9–46). New York: Academic Press.

Hardiman, P. T., Dufresne, R., & Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory & Cognition, 17,* 627–638.

Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In M. A. Just & P. Carpenter (Eds.), *Cognitive processes in comprehension.* Hillsdale, NJ: Erlbaum.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15,* 332–340.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13,* 295–355.

Homa, D. (1984). On the nature of categories. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 18, 1–47). New York: Academic Press.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Schwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology, 101,* 116–122.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 418–439.

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 322–330.

Keil, F. C. (1987). Conceptual development and category structure. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 175–200). Cambridge, England: Cambridge University Press.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Knapp, A. G., & Anderson, J. A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 616–637.

Lewis, M. W., & Anderson, J. R. (1985). Discrimination of operator schemata in problem solving: Learning from examples. *Cognitive Psychology, 17,* 26–65.

Mayer, R. E. (1975). Different problem-solving competencies established in learning computer programming with and without meaningful models. *Journal of Educational Psychology, 67,* 725–734.

Mayer, R. E. (1976). Some conditions of meaningful learning for computer programming: Advance organizers and subject control of frame order. *Journal of Educational Psychology, 68,* 143–150.

Mayer, R. E. (1983). Can you repeat that? Qualitative effects of repetition and advance organizers on learning from scientific prose. *Journal of Educational Psychology, 75,* 40–49.

Mayer, R. E., & Bromage, B. K. (1983). Different recall protocols for technical texts due to advance organizers. *Journal of Educational Psychology, 76,* 209–225.

Mayer, R. E., & Greeno, J. G. (1972). Structural differences between learning outcomes produced by different instructional methods. *Journal of Educational Psychology, 63,* 165–173.

Medin, D. L., & Ortony, A. (1989). Comments on Part 1: Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–196). Cambridge, England: Cambridge University Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 510–520.

Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 398–415.

Perfetto, G. A., Bransford, J. D., & Franks, J. J. (1983). Constraints on access in a problem solving context. *Memory & Cognition, 11,* 24–31.

Pirolli, P. L., & Anderson, J. R. (1984). Learning to program recursion. *Proceedings of the Sixth Annual Cognitive Science Meetings* (pp. 277–280), Boulder, CO.

Pirolli, P. L., & Anderson, J. R. (1985). The role of learning from examples in the acquisition of recursive programming skills. *Canadian Journal of Psychology, 39,* 240–272.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83,* 304–308.

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

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 629–639.

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, 15,* 456–468.

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.

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 5,* 484–494.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210,* 390–398.

Silver, E. A. (1979). Student perceptions of relatedness among mathematical verbal problems. *Journal for Research in Mathematics Education, 10,* 195–210.

Silver, E. A. (1981). Recall mathematical problem information: Solving related problems. *Journal for Research in Mathematics Education, 12,* 54–64.

Zhu, X., & Simon, H. A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction, 4,* 137–166.

Appendix A

Examples of Problems Used in the Experiments

Catch-Up Problems

Vat topic. Alexander is going to exhibit his exotic fish collection in an exotic fish show today. He has plenty of time to fill the aquarium. He decides to start filling it at 1:00 at the rate of 30 gal/hour to be ready by 4:00. At 1:00, he discovers, to his horror, that the show starts at 2:30. Alexander manages to get the tank filled on time. At what rate should he let the water flow into the aquarium?

Travel topic. Patty Pilot and her boyfriend Harold had a fight in Johnson airport. Patty got into her private plane and took off at 5:00, travelling at the speed of 180 mph. She landed in Millward airport at 10:00. Harold cooled off and decided to follow Patty. He boards a jumbo jet at 7:00. He lands in Millward at the same time Patty does, and they have another fight. How fast was the jumbo jet travelling?

Interest topic. Wella never invested money before and always kept her money in a checking account. Now, in June, she wants to invest in an account whose interest rate is high enough to allow her to earn 8%, on the average, for the whole year. At what rate must Wella invest her money to achieve this?

Work topic. Steve usually has no trouble filling his work quota by producing 30 widgets per hour during the usual 8-hour day. This morning, Steve showed up 3 hours late for work. The foreman wasn't too happy about this, but Steve promised to fill his quota by working much faster and not taking a lunchbreak. At what rate did Steve work for the rest of the day?

Dilution Problems

Vat topic. Sharon is making up her own organic garden fertilizer. The recipe says to drip ground seaweed into a vat of soy juice at the rate of 7 oz. per hour for 9 hours. Sharon got confused and set it up to drip at the rate of 9 oz. per hour. She realizes her mistake 4 hours later. She wants to turn the drip rate down and let the process continue for the full 9 hours. What should she turn the drip rate down to?

Travel topic. Susan is speeding along a toll road on a recent business trip. She travelled at 70 miles per hour for the first 4 hours of her 7½ hour trip. Then she remembered that she would have to pay any traffic violations out of her own pocket. She decided to slow down. When she reached the final toll gate, her average speed turned out to be 55 mph. What speed did Susan slow down to?

Interest topic. Tom invested part of the capital from his father-in-law's business in an account bearing 11% simple interest. Three months later, he discovered that the company's bylaws disallowed earning over 8% annually on investments. He manages to correct his mistake. He transfers the money to an account that will lower the company's average earnings for the year to 8%. At what rate should Tom reinvest the money?

Work topic. Eric's work team goes home early every day by making 8 circuit boards per hour for 6 hours instead of the prescribed 6 circuit boards per hour for 8 hours. Management becomes suspicious and places a spy on their team. The team discovers the plan on the spy's first day after 2 hours of work. They decide to slow their work rate down so that it will take them 8 hours to fill their quota. At what rate should the team work for the rest of the day?

Facilitation-Interference Problems

Vat topic. Chuck is developing a chocolate fondue recipe using a huge industrial vat. If he increases the rate of sugar flow by 5 gallons per hour, it takes 4 hours to mix in all the sugar. If he lowers the rate of sugar flow by 5 gallons per hour, it takes 6½ hours to mix in all the sugar. Since neither of the changes seems to make any difference Chuck gives up and goes back to the normal rate. What is the normal rate?

Travel topic. Jill, an aviation technician, is testing a model of an experimental jet plane in a wind tunnel. Flying against the air stream, it takes 10 minutes for the plane to travel the length of the tunnel. Flying with the airstream, it can travel the length in 6 minutes. Jill is amazed at how fast the plane can fly, especially since she knows that she set the wind speed in the tunnel to 15 mph. At what speed can the plane fly with the wind turned off?

Interest topic. At Christmas, Hilda gave each of two grandchildren an equal amount of money. One child put the money in a certificate. The other put the money in a bond. The rate on the certificate is 2% higher than the savings rate, and the rate on the bond is 2% lower than the savings rate. The certificate earned in 1 year the same amount as the bond did in 3. What is the savings rate?

Work topic. Sarah and John work at a bookstore packing crates. Sarah was recently hired, and John, having a crush on her, decides to help her out by secretly putting one of his packed crates in her output stack every hour. The management, seeing a decline in John's work rate, fires him. Sarah becomes despondent and packs one crate less per hour than she used to. Sarah now takes 8 hours to pack the same number of crates she used to pack in 5½ hours with John's help. What is her normal work rate?

Appendix B

Equations and Solution Procedures Used in Experiment 3

Category 1: Catch-Up Problems

$r_1t_1 = a_1 \quad r_2(t_1 - t_2) = a_2;\ t_2 < t_1$
If $a_1 = a_2$, then $r_1t_1 = r_2(t_1 - t_2)$

	Rate	×	Time	=	Amount
Goal rate	r_1		t_1		a_1
Higher rate	r_2		$t_1 - t_2$		a_2

Example:

$r_1 = 100 \quad t_1 = 35 \quad a_1 = ?$
But $a_1 = a_2$
$r_2 = \ ? \quad t_2 = 25 \quad a_2 = ?$

$r_1t_1 = r_2(t_1 - t_2)$
$100(35) = r_2(35 - 25)$
$3500 = r_2(10)$
$350 = r_2$

Check: $100(35) = 350(35 - 25)$
$3500 = 350(10)$
$3500 = 3500$

Category 2: Dilution Problems

$r_1t_1 = a_1 \quad r_2t_2 = a_2 \quad r_3(t_1 - t_2) = a_3;\ t_2 < t_1$
If $a_1 = a_2 + a_3$, then $r_1t_1 = r_2t_2 + r_3(t_1 - t_2)$

	Rate	×	Time	=	Amount
Goal rate	r_1		t_1		a_1
Higher rate	r_2		t_2		a_2
Lower rate	r_3		$t_1 - t_2$		a_3

Example:

$r_1 = 12 \quad t_1 = 20 \quad a_1 = ?$
$r_2 = 15 \quad t_2 = 10 \quad a_2 = ?$ But $a_1 = a_2 + a_3$
$r_3 = ? \quad t_3 = (t_1 - t_2) \quad a_3 = ?$

$r_1t_1 = r_2t_2 + r_3(t_1 - t_2)$
$12(20) = 15(10) + r_3(20 - 10)$
$240 = 150 + r_3(10)$
$90 = r_310$
$9 = r_3$

Check: $12(20) = 15(10) + 9(20 - 10)$
$240 = 150 + 9(10)$
$240 = 150 + 90$
$240 = 240$

Category 3: Facilitation-Interference Problems

$t_1(r_1 + r_2) = a_1 \quad t_2(r_1 - r_2) = a_2;\ t_1 < t_2$
If $a_1 = a_2$, then $t_1(r_1 + r_2) = t_2(r_1 - r_2)$

	Rate	×	Time	=	Amount
Standard rate	r_1		—		—
Composite higher rate	$r_1 + r_2$		t_1		a_1
Composite lower rate	$r_1 - r_2$		t_2		a_2

Example:

$r_1 = ? \quad t_1 = 3 \quad a_1 = ?$
But $a_1 = a_2$
$r_2 = 20 \quad t_2 = 6 \quad a_2 = ?$

$3(r_1 + 20) = 6(r_1 - 20)$
$3r_1 + 60 = 6r_1 - 120$
$3r_1 + 180 = 6r_1$
$180 = 3r_1$
$6 = r_1$

Check: $3(60 + 20) = 6(60 - 20)$
$3(80) = 6(40)$
$240 = 240$

Appendix C

Sample Orienting Task Questions Based on a Pair of Problems

Alexander is going to exhibit his exotic fish collection in an exotic fish show today. He has plenty of time to fill the aquarium. He decides to start filling it at 1:00 at the rate of 30 gal/hour to be ready by 4:00. At 1:00, he discovers, to his horror, that the show starts at 2:30. Alexander manages to get the tank filled on time. At what rate should he let the water flow into the aquarium?

Wella never invested money before and always kept her money in a checking account. Now, in June, she wants to invest in an account whose interest rate is high enough to allow her to earn 8%, on the average, for the whole year. At what rate must Wella invest her money to achieve this?

Recognition Questions

1. He decides to start filling it at 1:00 at the rate of 30 gal/hour to be ready by 4:00.
2. He plans to start filling it at 1:00 at the rate of 30 gal/hour to be ready by 4:00.

1. At 1:00, he discovers, to his horror, that the show starts at 2:30.
2. At 1:00, he discovers, to his horror, that the show will begin at 2:30.

1. Now, in June, she wants to invest in an account whose interest rate is high enough to allow her to earn 8%, on the average, for the whole year.
2. Now, in June, she wants to deposit money in an account whose interest rate is high enough to allow her to earn 8%, on the average, for the whole year.

1. Wella never invested money before and always kept her money in a checking account.
2. Wella never invested money before and always deposited her money in a checking account.

Verification Questions

1. Alexander plans to start filling the tub at 2:30.
2. Alexander plans to start filling the tub at 1:00.

1. Alexander actually starts filling the tub at 1:00.
2. Alexander actually starts filling the tub at 12:00.

1. Wella wants to earn 8% simple interest annually.
2. Wella wants to pay out 8% simple interest annually.

1. Wella earned interest on her money from January to June.
2. Wella earned no interest on her money from January to June.

Analogy Questions

30 gal/hour : filling tank from 1:00 to 4:00 :: 8% : earning interest from ?
1. January to June
2. January to December

Filling tank from 2:30 to 4:00 :: earning interest from ?
1. June to December
2. January to December

Fill : aquarium :: ? : investment account
1. earning interest
2. withdrawing interest

Filling tank from 1:00 to 4:00 :: earning interest from ?
1. January to December
2. January to June

Schema Questions

1. Both problems involve a situation where a lower rate has to be computed in order to lower the average overall rate at which something was accomplished.
2. Both problems involve a "catch-up" situation, where there is less time than was anticipated to achieve some goal and a higher rate must be computed.

[Note: These schema questions describe the dilution and catch-up problem structures, respectively. The schema question for the facilitation-interference problem structure was as follows: "Both problems describe a situation where a standard rate is increased and decreased by a constant."]

Received September 18, 1991
Revision received March 2, 1992
Accepted March 11, 1992 ■