Journal of Experimental Psychology:
Learning, Memory, and Cognition
1996, Vol. 22, No. 3, 792-810

0278-7393/96/$3.00

Content Effects in Problem Categorization and Problem Solving

Stephen B. Blessing and Brian H. Ross
University of Illinois

In many domains, the content of a problem (i.e., its surface cover story) provides useful clues as to the type of problem it is and to its solution. Five experiments examined this role of problem content on the problem solution and categorization of algebra word problems with experienced participants. In the first experiment, when problem content was atypical for the problem's deep structure, people were worse at solving the problem. Differences were also detected in the problem solution where the problem's content was highly correlated with its deep structure versus problems where content was neutral to their deep structure. In the other experiments, problem categorization and determination of information relevance depended on how closely the problem's content matched its deep structure. These results suggest that content may be influential even for experienced problem solvers. The discussion examines the implications for problem schema access and application.

The goal of this article is to examine how the specific content of a problem may affect the problem solving of experienced solvers. In most domains, there is an empirical correlation between problem types and problem contents. We argue that experienced problem solvers learn to make use of these formally irrelevant, but empirically predictive, contents in accessing and applying their relevant knowledge. Understanding these content effects is important not only because content affects performance, but also because these effects provide clues as to how experienced solvers represent and use this relevant knowledge. We first provide some background on problem content and expertise and then return to this issue.

Problem Content and Expertise

Word problems are common in many formal domains. Many researchers have made the distinction between a word problem's surface structure and deep structure (e.g., Chi, Feltovich, & Glaser, 1981). The surface structure includes the settings, events, and objects mentioned in the problem. We refer to this as *content* or *surface content.* A problem's surface structure is distinct from its deep structure, which is the set of principles or equations important for solving the problem. For example, consider the following algebra problem: "A riverboat travels 30 miles downstream going with the current. In an equal amount of time the riverboat travels only 20 miles upstream going against the current. The riverboat is capable of going 5 mph when there is no current. What is the rate of the current?" The boat, river, and the river's current, along with the description of the physical events, constitute the example problem's content. To solve this problem, a person would use the formula distance equals rate times time (with the rate modified by the river's current), and so that is the problem's deep structure.

Why is this distinction between content and deep structure important? In their seminal article, Chi et al. (1981) found that novices and experts differentially relied on the surface content and deep structures for categorizing physics problems. Novices categorize physics problems on the basis of surface content. For example, beginning physics students will often categorize problems based on whether they include inclined planes, springs, or pulleys (see also Schoenfeld & Hermann, 1982, and Silver, 1979, for similar results in other domains). Furthermore, the knowledge that the novice brings to bear in solving these problems is closely tied to these content features (e.g., Bassok, 1990; Bassok & Holyoak, 1989; Novick, 1988; Ross, 1984). Experts, however, use a problem's deep structure for categorizing and solving problems. This distinction between relying on surface content versus deep structure is often considered a primary difference between novices and experts (e.g., Reimann & Chi, 1989). Indeed, Schoenfeld and Hermann's results provide good evidence for this point. Their participants, novices in mathematical problem solving, initially categorized such problems on the basis of surface content. However, after taking an intensive course in mathematical problem solving (presumably becoming more expert in this task), they based their categorizations on the problems' deep structure.

Despite the findings that experts often focus on a problem's deep structure, we believe that experts often do make use of a problem's content during problem solving. Although the content of a problem may seem irrelevant to its solution, a notion

Stephen B. Blessing and Brian H. Ross, Department of Psychology, University of Illinois.

This research was supported by Air Force Office of Scientific Research Grant 89-0447. Experiments 1 and 2 were conducted by the first author while at the University of Illinois as part of his honors thesis work under the guidance of the second author. Experiment 3 was conducted at the Beckman Institute for Advanced Science and Technology. Parts of these experiments were presented at the Sixteenth Annual Conference of the Cognitive Science Society, Atlanta, Georgia, August 1994. We thank the many colleagues who have provided comments on the work and manuscript, including Denise Cummins, John Hayes, Matthew Kilbane, Marsha Lovett, Gregory Murphy, Colleen Seifert, Thomas Spalding, and especially John Best. We would also like to thank Tonya Sieverding for her help in scoring the data.

Correspondence concerning this article should be addressed to Stephen B. Blessing, who is now at Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890. Electronic mail may be sent via Internet to blessing+@cmu.edu.

that the research discussed above would support, it is often the case that a particular word problem type has a "typical" content. That is, a class of problems with an almost identical deep structure will usually share a similar content. For example, in physics it is often the case that in solving a problem with an inclined plane, Newton's Second Law will be needed. In the algebra problem mentioned previously, problems about riverboats almost always use the same modified distance equals rate times time equation. Mayer (1981) conducted an analysis of algebra textbooks and found a very strong relation between the deep structure and content of algebra word problems. With this correlation between contents and problem types, it would be strange if experts did not take advantage of content to access the appropriate knowledge. Thus, understanding the content effects in these domains, such as the algebra domain studied here or even in less formal domains, is a necessary part of a theory of experienced problem solving.

Evidence for Content Effects in Experienced Problem Solvers

There is some evidence that experts can and do make use of the surface content of the problem in some situations. In one of Chi et al.'s (1981) experiments, they reported that experts used keywords from the problem for activating potentially relevant knowledge and that these keywords were a subset of the ones used by novices. Furthermore, Hardiman, Dufresne, and Mestre (1989) found that experts' decisions as to what physics problems were similar to one another were affected not only by the problems' deep structure, but also by the similarity of the problems' contents. Novick (1988) showed similar results within a problem-solving task, in which experienced problem solvers were initially misguided by the surface contents of problems that were similar to the contents of earlier problems. These three studies each indicate the potential importance that experts place on a problem's content. As stated previously, because a problem's deep structure is often correlated with content, a reliance on content is often not a mistake.

Most directly relevant to the current work is the influential article by Hinsley, Hayes, and Simon (1977). Their study provides insight into the importance of content to people experienced in the domain. They showed that people experienced in algebra can categorize algebra word problems into certain, definite types (e.g., interest, river current, mixture, work) and can do so soon after beginning to read the problems. For example, after hearing only the starting noun phrase "A river steamer . . . ," people who are familiar with such river current problems can give an adequate description of the rest of the problem, and also information concerning how the problem would be solved (e.g., Mayer, 1982). Their experienced algebra problem solvers used such category information to start formulating the problem for solution, before having read the entire problem. Even if the problem consisted mostly of nonsense words, the solvers would attempt to place the problem into a category, presumably because it aids the problem-solving process. Hinsley et al. also provided some preliminary evidence that a problem's content can affect the solution method with which a person will try to solve a problem. Verbal protocols indicated that the problem solvers were more likely to implement a previously acquired solution procedure (a schema) to solve problems that contained typical contents. When the problem's content was atypical for the problem's deep structure, people more often used a line-by-line solution method. Further details of these experiments are discussed later as motivations for particular experiments.

Problem Schemata

Why might even experts come to rely on content as a means of categorizing and solving problems? A common way of characterizing problem-solving knowledge is in terms of problem schemata, adapted from the general idea of schemata (Rumelhart & Ortony, 1977). Problem schemata are knowledge structures that are used to identify the type of problem being solved and that contain associated procedures for solving problems of that type. The relation of the problem to the relevant schema may be both bottom-up and top-down, in that the problem information, such as surface content, helps to access the schema and then the schema information aids in instantiating the rest of the problem. One of the crucial research issues in problem solving is to understand problem schemata: how they are represented, acquired, accessed, and applied (e.g., Reimann & Chi, 1989).

Thinking in terms of problem-solving schemata provides two further motivations for why content might be expected to influence problem solving. We mention these ideas briefly here, but both are elaborated in the General Discussion. First, much of the work investigating analogical problem solving shows that novices are often reminded of prior problems on the basis of the content of the present problem, not its deep structure (e.g., Holyoak & Koh, 1987; Ross, 1984). Although our current work is examining the access and application of schemata and not of specific prior problems, the processes for the two may be similar, because there is likely to be a gradual increase in the generalization of knowledge being accessed during problem solving. The surface contents provide means by which novices can begin to compare problems of a given type and form generalizations of that problem type (e.g., Ross & Kennedy, 1990). Furthermore, recent work by Bernardo (1994) shows that these generalizations will often include aspects of specific problems. To the extent that problems of the same type share a similar content, it is likely that such content information may never be abstracted away within an expert's schemata, and so experts may still use content in helping to solve problems.

Second, if schemata are to be accessed early during problem solving so that they can be used to interpret the rest of the problem, then the use of content would be greatly advantageous. As mentioned earlier, if the surface contents and deep structures are empirically correlated, then the experienced solver can use the surface contents to access the likely problem schema. The work of Hinsley et al. (1977) was motivated by this idea, and their results indicate that content that is predictive of the problem type is used to access the problem schema.

Overview of Experiments

Most of the research with experienced problem solvers has examined how they make use of the deep structures of the problems. Although the above research implicates the potential importance of problem content for experienced problem solvers, little research has directly examined these content effects. It may be the case, however, that although experienced people are able to use content in solving a problem, a difference may not be detected in problem-solving performance when content is varied but when the deep structure remains the same. That is, even though they may be able to use content to access a schema, they also have other methods of categorizing and solving the problems, like the problem's deep structure. Problems with different contents but the same deep structure may be approached in a similar way by the experienced problem solver. Cummins (1992) argued that the reliance on surface content versus deep structure of a problem may be more properly viewed as a difference in weighting, rather than a qualitative difference. As a person becomes more expert in a task, they shift from placing more importance on content features to placing more importance on deep structure.

However, given Hinsley et al.'s (1977) results, content may still play an important role in expert problem solving. Whereas novices may use content in a very superficial manner, experienced solvers may use content to access and apply relevant information for solving the problem. In these experiments, we investigated how experienced problem solvers make use of content while solving problems. As we have argued earlier, these potential effects of content are of interest both because of their effects on performance and because they may be useful in better understanding how such experienced solvers represent, access, and use their relevant knowledge. In many respects, this current work can be viewed as an extension of Hinsley et al.'s study described earlier. Their experiments provided some suggestive results on the potential importance of problem content in categorization and problem solving. However, their purpose was to examine evidence for the existence of problem schemata, so content manipulations were incidental.

We first gathered some basic performance data on the effects of content in problem solving. (In many of the previous studies examining the effect of content on these tasks, the researchers only examined a few participants in detail.) Our working hypothesis was that the content of the problem will affect problem-solving performance. In particular, we assumed that experienced solvers can make use of the empirical correlations existing between contents and problem types. If so, then relative to a baseline with neutral contents, problems for which contents typically occur with a given type should be easier to solve, but problems for which contents mismatch the problem type should be more difficult to solve. These performance data are supplemented with problem-solving protocols to provide a fuller picture of the effects.

In addition, we provide a more detailed analysis of these content effects. In particular, we investigated two important ways in which the content effects may occur, based on the general analysis of how problem schemata may influence performance. First, schemata allow rapid categorization of problems into their types (to allow the application of relevant knowledge). Problem categorization is often considered to be a crucial step in problem solving (e.g., Chi et al., 1981). In Experiment 2, we examined how content affects problem categorization. Second, schemata, by a combination of inference and top-down processing, allow problem solvers to determine what problem information is relevant to the solution and what information is irrelevant. In Experiments 3a and 3b, we examined the effect of content on the determination of problem information relevance. The results of these experiments provide both fundamental data on content effects and an initial investigation of how content is incorporated into the relevant problem-solving knowledge of experienced problem solvers.

Experiments 1a and 1b

We began by examining whether differences in the content of algebra word problems affect basic performance measures (accuracy and latency). In addition, we supplemented these findings by collecting verbal protocols from additional participants. The protocol results will be discussed in Experiment 1b.

Experiment 1a

For these experiments, we wrote triplets of algebra word problems that differed in their surface content but had the same deep structure. One problem in the triplet had a content that was *appropriate* for that triplet's deep structure. For example, the problem might look at different investments gaining interest at different rates (for an interest problem) or a boat going up and downstream (for a river current problem). Another problem in the triplet had a content that was *neutral* to the deep structure (e.g., using flower blooming for an interest problem or going up and down escalators for a river current problem). The last problem in the triplet had a content that was *inappropriate* for the deep structure, using the content of a different problem type (e.g., using a content about a taxi driver for the interest problem or using a content about people working for the river current problem). More on the structure of these problems is given in the *Materials* section.

On the basis of the work reviewed in the introduction, we expected that problem content would affect problem solving. In particular, the prediction was that the appropriate condition would lead to better problem solving performance than would the neutral condition and that the inappropriate condition would lead to worse performance. For example, Novick (1988) found that inappropriate content can initially mislead experienced problem solvers, though her content manipulation was in terms of similarity to a previous problem, not to typical contents. Hinsley et al. (1977) did compare what we would call appropriate versus inappropriate conditions, finding a difference in the method of solution, but tested only 2 participants (graduate students) on nine problems each. In addition, without a neutral condition, it is difficult to know if the content effects were facilitative, interfering, or both. Thus, our work extends the earlier findings plus provides a basis for understanding the further extensions in the later experiments.

Method

Participants. The people tested in this experiment were 24 graduates of the Illinois Mathematics and Science Academy (IMSA) who were attending the University of Illinois at the time of the experiment. IMSA is a residential high school in Aurora, Illinois for academically gifted students. These participants were used to ensure a high degree of math competency and previous math experience. Although perhaps not algebra experts, they can be considered highly experienced in the subject. All of them had had math through calculus, and many had participated on the school's math team, whose competitions often feature word problems. They were paid $5 for their participation in the experiment, which lasted about 45 min.

Materials. We constructed triplets of algebra word problems that had the same deep structure but different contents. That is, a set of problems was written that are solved in the same way and have the same amount, kind, and placement of information, but contain different cover stories. The first three problems in Appendix A all use the underlying structure found in age problems, which generally deal with the relation between people's ages. The appropriate problems have the content typically associated with the problem type (i.e., the content correlated most strongly with the problem type), such as the ages of Michelle and her niece in the first problem of Appendix A. The neutral problems have a content that is not usually associated with any particular problem type, such as the relation between exam scores in the second problem of Appendix A. Finally, the inappropriate problems have a content typically associated with a different problem type, such as using the content for a work problem for the third problem in Appendix A. As much as was possible, we tried to write the problems so that the only way in which the matched problems differed was in the nouns (and related verbs) used in the problems.

Because the ways in which the types are defined and the content is manipulated are central to these experiments, a fuller explanation of each is required. The six problem types were chosen from the ones identified by Hinsley et al. (1977) and are commonly used in algebra textbooks: age, interest, mixture, motion (distance equals rate times time), river current, and work. Although these different problem types are labeled in terms of their typical contents, they generally differ in terms of their deep structure as well. Mayer (1981) identified eight separate families (which differ in their deep structure) for algebra word problems. Within each family are differing numbers of categories, which are the usual level of analysis in algebra textbooks and are equivalent to our problem types. Our six problem types come from four different families (time rate, percent cost rate, straight rate, and number story). Three problem types came from one family (time rate), but different categories (motion, current, work).[1] It could be the case that content effects are observed for only a subset of the problem types. To ensure that our results are not due to a failure to adequately sample from differing deep structures, we will supplement the overall findings with the results for the three types from different families and the three types from the same (time rate) family. Table 1 provides more information about the deep structures that we used with regard to Mayer's (1981) taxonomy. The prototypical equations used to solve each problem type are given underneath the problem type name. The constants A, B, C, D are often given in the problem statement, though not always, with the usual goal to solve for either x or y. The equations for the mixture and interest problems are similar, but they are still distinct from one another, and Mayer places them in different families.

The typical content for each problem type from Mayer (1981) is also given in Table 1. Our manipulation of content for the three conditions was almost perfectly consistent with these (the one exception being our more specific mixture contents). For the appropriate condition, the problem used the typical contents of Mayer's taxanomy, with elaborations, such as the ages of a niece and aunt for the age problems. For the inappropriate condition, the typical content (with elaboration) of the inappropriate problem type was used. For the neutral content, the contents we chose are ones that do not fall within the typical contents given by Mayer (with the exception of the mixture type). In most cases this led to a problem that was rather different than the usual problem for the problem type, such as the rates at which people caught fish (work), the number of acorns collected by animals over different days (age), or the rate at which different colored rabbits produced offspring (interest). In a few cases, this manipulation may not appear to fully remove the problem from the appropriate condition (e.g., one reviewer pointed out that the arrow problem in the neutral motion problem presented in Appendix A seems by content to be a motion problem). We are not sure how to objectively define the neutral manipulation other than to argue that the neutral condition does not use the typical objects given by Mayer. Additional examples are given in Appendix A.

Two appropriate problems of each of these six types were constructed, for a total of 12 problems. Twelve neutral and 12 inappropriate content problems were matched to this appropriate set. We refer to these 12 triplets (the matched appropriate, neutral, and inappropriate problems) as the 12 base problems. The number of words and syllables, the numbers used, and the placement of information were very similar among the three conditions. The equations were exactly the same across conditions, though in some instances the numbers were changed by a factor of 10 to better fit with the content. Although when one changes content it is not possible to rule out all possible differences, the materials were matched as closely as possible. Summary statistics for the three conditions are available from us.

These 36 problems were split into two sets of 18 problems, with both sets having each of the six problem types in the three conditions. Participants received problems of only one of the two sets, and the problems used in all experiments were counterbalanced so that every problem had the same number of observations. In this experiment, every participant received 12 of the 18 problems from a set, but the problems were chosen so that every problem type (e.g., river current, interest) was represented twice and every condition (appropriate, neutral, and inappropriate) was represented four times. (Because both the neutral and inappropriate content problems were based on the original 12 appropriate content problems, participants could only receive a maximum of 12 problems in order to avoid duplicates.)

Procedure. Participants were tested in small groups of 2–6. They were given a booklet containing 15 problems: 3 practice problems and 12 test problems. The problems were presented to the participants 1 per page, with the problem printed at the top. The 3 practice problems were of different types (perimeter, probability, and Pythagorean theorem) than the six types listed previously, but were similar in terms of difficulty. The 12 test problems contained two of the six types, with 4 in each of the three conditions. The problems were randomly ordered for each participant, but participants received 1 problem of the six types before receiving another problem of a previously seen type. Furthermore, as a result of their similarity, interest and mixture problems were never presented one after the other. Every problem was presented to 8 participants.

After the experimenter read the instructions aloud, the participants turned to the first problem and began work. After every 45 s, the experimenter would call out, "Line." The participants then drew a line across the page below where they had been working and continued work below the line. In this way, a rough estimate of the amount of time participants spent solving every problem was obtained. The

[1] We were not aware of Mayer's (1981) classification when we first constructed the materials. However, as will be shown with each experiment, the pattern of results is the same for the three problem types from different families and the three problem types from the same category.

participants had 3 min to work on every problem, and they were not permitted to work ahead.

Results

The dependent measures were the accuracy and latency with which the participants solved the problems. The accuracy score for the problems was a score between 0 and 1. A score of 1 was given if the solution was correct, and a score of .75 was given if a small mistake, such as a mathematical error, was made. A .5 was awarded to solutions that contained a conceptual error (such as reversing correspondences), and a .25 was awarded to solutions with more than one error, but for which it still appeared that the participant had some correct notion of how to solve the problem. Erroneous and missing answers received no credit. The problems were scored by two scorers, and the discrepancies were adjudicated by a third party.

Participants scored a mean of .73 on the appropriate content problems, .77 on the neutral content problems, and .64 on the inappropriate content problems, $F(2, 23) = 3.76$, $MSE = 0.41$, $p < .05$. A Newman-Keuls test showed that the inappropriate content problems differed from both the appropriate and neutral problems ($p < .05$). Other scoring schemes (e.g., no partial credit) yielded the same pattern of results. The inappropriate condition performance was the worst for 16 of the 24 participants and for 6 of the 12 base problems. We further analyzed the poor performance in the inappropriate condition to see if it was due to poorer performing participants or particular families of problems. We did not find any clear relations: The 16 participants whose performance was worst on the inappropriate problems did not have lower overall accuracy scores than the remaining eight participants (.69 vs. .67). The six inappropriate problems that led to worse performance came from four different problem types (age, interest, motion,

Table 1
Information on Problem Types

		Mayer's (1981) taxonomy		
Problem type	Equation	Family	Templates used	Typical content
Age	$x = Ay$ $x + B = C(x + B)$	Number story	Relative now then	People's ages at two different times.
Interest	$x + y = A$ $Bx + Cy = D$ (where B and C are percentages)	Percent cost rate	Two absolute amounts Two relative amounts	Money is split and gains interest at two rates.
Mixture	$x + y = A$ $Bx + Cy = AD$	Straight rate	Two absolute amounts	Two solutions are mixed to yield a new solution.
Motion	$y = Ax$ $y = B(x + C)$	Time rate	Equal time Overtake	Moving vehicles, either travelling to the same point in an equal time or one is trying to overtake the other.
River current	$y = (x - A)B$ $y = (x + A)B$	Time rate	Round trip relative Total time	A boat moving with or against a current.
Work	$y = Ax + Bx$ (where A and B are rates)	Time rate	Together absolute	Two people who work at different rates trying to complete a shared task.

Note. For more information on the problem families and templates, refer to Mayer (1981).

and work). Participants performed similarly across the three conditions on the two remaining problem types (mixture and river current). The appropriate and neutral conditions showed no clear difference, with the appropriate leading to better performance than the neutral for 8 of 20 participants (four ties) and 3 of 10 base problems (two ties).

As previously mentioned, latency was assessed by having the participants draw lines across the page at 45-s intervals. Because participants had 3 min to solve the problems, there were four 45-s intervals for each problem. The midpoint of the interval in which the participant wrote down the equation that once solved would yield their final answer was recorded for that problem. A score of 3 min was given if the participant never arrived at such an equation. Latency was measured in this way rather than simply noting the finishing time, because we were interested in the time-to-equation and did not want to include the additional time for arithmetic operations and checking of answers, which could differ considerably among problems. This method, although only giving gross measures, intrudes only a little on the participant's solutions and provides a running time measure. The participant could perhaps arrive at an incorrect equation that was used to find their, probably incorrect, answer. The time measure for that problem would be the interval during which the participant wrote down that incorrect equation. (The time to correct equation, given later, shows similar results, but we felt that the time to any equation was more likely to provide a latency measure that did not depend greatly on the accuracy score.)

Participants spent a mean of 1.06 min to write the equation on the appropriate content problems, 1.19 min on the neutral content problems, and 1.63 min on the inappropriate content problems, $F(2, 23) = 15.38$, $MSE = 3.95$, $p < .01$. A subsequent Newman-Keuls test again showed that the inappropriate content problems differed from both the appropriate

Content in the present study		
Appropriate	Neutral	Inappropriate
People's ages at two different times.	Two squirrels collecting acorns on different days.	A mason mixing cement in two different containers, and then adding some.
	Comparing two people's test scores before and after the teacher adds points.	A chemist mixing chemicals into two flasks and then adding some.
An investor receives dividends from two different accounts.	A person plants flower bulbs that reproduce at different rates.	A taxi (or limo) driver drives two different cabs (or limos) a different percentage of some total.
	A rabbit breeder has rabbits that reproduce at different rates.	
A chemist mixes two acids to make one solution of a certain acidity.	A person at a party mixes two types of punch to make a single drink.	A person goes to two birthday parties where people are of different ages.
A metallurgist mixes two alloys to obtain an alloy with a specific purity.	A car company has two factories that produce different percentages of red cars.	A birthday club has people with two different ages in it.
Two drivers, both go to the same destination.	Two archers shooting arrows at a target.	Two investors, both trying to reach $1000.
Two trains, one starts later but goes faster.	Two football players, one trying to catch the other.	Two investors invest money at different rates.
A boat goes a certain distance with the stream relative to against.	A trolley goes a certain distance down a hill relative to up the hill.	Two people working together, but at different relative rates.
A boat goes a certain distance with the stream and against the stream.	A boy goes a certain distance walking either with or against the down escalator.	One person that has 0, 1, or 2 helpers doing a certain task.
Two people who work at different rates on a common goal.	Two types of fish who breed at different rates trying to fill pond.	Two tugboats that operate at different rates, trying to move one boat.
	Two fishermen who fish at different rates trying to catch so many fish.	Two engines on one boat that operate at different rates.

and neutral content problems ($p < .05$). Eighteen of the 24 participants took longest to solve the inappropriate content problems, and it was slowest for 8 of the 12 base problems. Again, there was no difference between the appropriate and neutral conditions, with the appropriate performance higher for 12 of the 24 participants and 6 of the 12 base problems.[2]

These results are buttressed by an additional study that used the same design, but tested 24 students from the University of Illinois psychology participant pool. Within this population of students, we found the same pattern of results, but poorer performance overall. They achieved a mean accuracy of .36 on the appropriate content problems, .38 on the neutral content problems, and .21 on the inappropriate content problems, with corresponding latencies of 2.05 min, 2.01 min, and 2.34 min. Although these students may have not had quite as much experience (or ability) as the IMSA students, they had had quite a bit of experience in solving algebra word problems. (All entering students to the University of Illinois School of Liberal Arts and Sciences must have at least 3.5 years of high school math, which would typically include 2 years of algebra.) They appeared to know many of the problem types, but were unable to remember well how to go about solving each of the types. Thus, we think it is probably most reasonable to think of these University of Illinois students as ones who had been experienced algebra word problem solvers, but were suffering from not having worked on such problems for a few years.

One last analysis examined time to correct equation. For this analysis, only problems that received an accuracy score of 1 or .75 were used. There were 76 appropriate problems that met that criterion, and participants spent a mean of 0.91 min solving them, compared with 83 neutral content problems at a mean of 1.05 min, and 70 inappropriate content problems at a mean of 1.42 min. Even though participants were faster overall doing problems they solved correctly, inappropriate content problems still took longer to solve.

Discussion

The results did show an effect of content on the problem solving performance of experienced problem solvers, but the effect was not exactly as predicted. We focus here on two main results. First, as predicted, the inappropriate condition performance was less accurate and had longer latencies than the performance in the other two conditions. This finding shows that there was an effect of content on problem solving performance. Novick (1988) found that content similarities from within the experiment can misguide even experienced solvers, and the results here extend this finding to preexperimental content effects. Using the schema framework, it may be that the inappropriate content activates an inappropriate schema, misleading the problem solver. We examine this idea further in the problem-solving protocols and Experiment 2.

Second, contrary to our prediction, the appropriate and neutral conditions did not differ in either accuracy or speed. The lack of any difference even in latency was surprising, but the dependent measure used was a gross one (because we assumed there would be accuracy differences) and it may be that the difference in categorization, or schema access, times would be small anyway. In addition, because the neutral contents simply used nontypical objects, it may be that it was not a strong enough manipulation. In the following experiments, we further investigated this lack of difference. A final comment to make about this lack of appropriate–neutral difference is that the problems were generally easy ones and it may be that more difficult problems are required before a difference emerges (see Experiments 3a and 3b).

Experiment 1b

To examine more qualitatively the effects (and lack of effects) from Experiment 1a, a follow-up study was conducted in which participants were instructed to think aloud while solving the problems. We examined these protocols for the type of solution method experienced problem solvers used for the problems within the three content conditions. Hinsley et al. (1977) performed a similar type of experiment. In that experiment, they had three problem types crossed with the three contents normally associated with those three types. This resulted in nine problems, three appropriate and six inappropriate content problems. They found, based on testing 2 people, that the appropriate problems were often solved by a schema method while the inappropriate problems were often solved by a sentence-to-equation method. We expected to see a similar pattern with our appropriate and inappropriate problems. Furthermore, we hoped that from an analysis of the protocols we could better understand the lack of difference found between the appropriate and neutral problems in the main experiment. If it is the case that most of the neutral problems are solved with a schema-based method, then perhaps solvers are able to figure out the problem's type through some means other than content. However, if people solve the neutral problems using the sentence-to-equation method, then solvers do not access category information to solve the problem but, as we saw in Experiment 1a, are still solving the problems at a level equal to the appropriate content problems.

Method

Six additional participants were tested, using the same three practice problems and six experimental problems (one of each problem type, two in each of the three conditions. Each of the 36 problems was solved by 1 participant). Prior to the experiment, the participants

[2] There could be an effect of problem order, because participants received two of each problem type. To examine this, we analyzed only the first six problems, which include only one of each type, that the participants solved. The results show the same pattern as the overall data. For the accuracy data, the results were .75 for appropriate content problems, .78 for neutral content problems, and .60 for inappropriate content problems, $F(2, 23) = 3.76$, $MSE = 0.228$, $p < .05$. For the latency data, the results were 1.04 min for appropriate content problems, 1.04 min for neutral content problems, and 1.86 min for inappropriate content problems, $F(2, 23) = 13.33$, $MSE = 2.82$, $p < .01$.

In addition, the three problem types from three different families showed the same pattern of performance as the three types from the same (time rate) family. For the different families, performance was .74, .79, and .64 for accuracy and 0.91, 0.96, and 1.50 for latency. For problems from the same family, the corresponding numbers were .71, .75, and .64, and 1.23, 1.41, and 1.77.

were instructed on how to give a think-aloud protocol, and were given a warm-up task of adding numbers together while thinking aloud. They read the problem aloud and were asked to think aloud while solving the problems. They had pencil and paper to write out the solutions and the protocols were taped for later transcription.

Results and Discussion

Both accuracy and time to equation were again measured for the three conditions. Using the same scoring criteria as before, participants' mean score was .85 for appropriate content problems, .75 for neutral content problems, and .79 for inappropriate content problems. Participants spent an average of 1.33 min producing their main equation for appropriate content problems, 1.65 min for the neutral content problems, and 2.17 min for the inappropriate content problems. Although the accuracy scores do not show the same pattern as Experiment 1a and the follow-up study mentioned, $F(2, 5) = 1.72, MSE = 0.001, p > .1$, the latency measure does, $F(2, 5) = 23.4, MSE = 0.04, p < .01$, with the inappropriate problems solved significantly slower than in the other two content conditions, suggesting that the inappropriate problems were particularly difficult. The neutral condition did lead to both worse and slower performance than the appropriate condition, unlike the earlier findings, but the differences were not large, given the variability arising from the small number of observations. For collecting protocols, the experimenter tested each participant individually and without a time constraint, perhaps leading to less difference in accuracy among the conditions. However, the latency pattern was replicated from the main experiment (and latency was measured more accurately than in the main experiment because the latencies were recorded from the tapes). Given that we obtained the same accuracy pattern with both IMSA and University of Illinois students in experiments with more observations, it seems likely that the protocol accuracy results may have been due to greater noise with smaller numbers of observations.

Of more interest here, however, are the methods used by the participants when solving these problems, as evidenced by their think-aloud protocols. We used the following scoring criteria: A sentence-to-equation solution is one where the participant takes a sentence in the problem statement and translates it into algebra. Once two or three sentences are translated, the problem can generally be solved. A schema-like solution occurs when the participant can go directly from reading the problem statement to writing down the correct equation. That is, the participant immediately writes down the one needed equation to solve the problem, implicitly using any intermediate equations that might be needed by substituting them into their one equation. Participants using a schema usually do so immediately after reading the problem. Participants who do not immediately start using a schema will begin to translate the sentences into equations and end by solving the problem that way. Table 2 provides an example of both solution types. The participant who solved the appropriate work problem almost immediately after reading the problem statement worked out the right answer in the way illustrated. In the sentence-to-equation solution, however, the participant who solved the inappropriate version of the same problem did not solve the problem in the same way. Rather, after initially trying a different solution (one more consistent with a river current problem), the participant proceeded, as illustrated, in a relatively stepwise fashion in finding the final solution.

Table 2
Sample Protocols for Schema-Like Versus Sentence-to-Equation Solutions in Experiment 1b

Schema-like (appropriate work content for work problem)
OK, well, so the apprentice can do a job in half the time that the electrician. I am . . . In twice the time the electrician can so he is like half an electrician. So basically what you have is an electrician and a half. So that would be 2 divided by 3 halves or 2 times 2/3 or 4/3 of an hour.
Sentence-to-equation (inappropriate river current content for work problem)
If Ship A can go 10 miles in 2 hours equals 5 mph, if the other one takes 4 hours, it goes 10 miles in 4 hours, or 10 miles for 4 hours is 5/2 mph. How long does it take together if they push together? So then you want to know this one in half miles . . . 10 halves. So, 10 plus 5, 15 halves, and you need to go 20 halves. They need 5 halves more, how long would 5 halves take? 5 is a third of 15, so another one-third of an hour. So at the rate of 15 halves, it would take another hour and 20 minutes.

Note. Refer to Appendix A for the problem statement.

In doing the analysis presented below, we independently examined the protocols and classified the solutions as either schema-like or sentence-to-equation and then settled the few differences. We also had a scorer blind to condition rate the solutions for their method. Her ratings coincided with ours on 35 of the 36 observations (.97).

For almost all appropriate content problems (10 of 12), the participants had a schema-like solution, which replicates Hinsley et al.'s (1977) findings. That is, the solvers seemed to recognize the problem type, often stating a category name while reading the problem statement, and then applying stored knowledge to solve the problem. Participants were able to go directly from reading the problem to setting up the necessary equation. For example, 1 participant midway through reading a problem about trains said, "I always hate these rate problems," and then immediately set up the equation $d = r \times t$ (distance equals rate times time), the needed equation to solve the problem.

The method used to solve neutral content problems seemed to be split evenly (6 and 6) among using a schema and the sentence-to-equation method. Promptly after reading an interest problem involving rabbits breeding (rather than the typical content about money and banks), 1 participant was able to formulate the problem in one step and then quickly solve it. However, 1 participant after reflecting on an age problem about squirrels collecting acorns, wrote down a four-equation, three-unknown system of equations and proceeded to laboriously solve the problem. Usually these problems are solved using only two equations and one unknown.

For the inappropriate problems, some participants used a schema (5 cases out of 12). It appeared that on these problems, the participant was able to ignore the content and use other clues. For example, mixture problems are usually about a chemist mixing two liquids of different concentration together

to obtain one bottle of liquid with a new concentration. To solve these, one must usually use an equation involving an elaboration of a weighted average (i.e., with the final concentration as a weighted average of the two initial concentrations, weighted by the amounts of liquid). In writing the problems for the inappropriate condition, we put the word *average* in the problems (whose content concerned birthday parties and different ages; see Appendix A), and the participant used that word cue. However, on 4 of the 12 inappropriate problems, participants incorrectly classified a problem early in reading and that incorrect classification adversely affected their problem-solving performance. As an example, 1 participant read the beginning of an interest problem with a train content of a motion problem and said, "Hey, I think of bullet trains hitting each other coast to coast and you want to find out exactly when." Such a set up and question is often the case in motion problems, but was not helpful in solving the underlying interest problem type. Miscategorizations never occurred for the neutral or appropriate problems. However, participants could recover from an initial incorrect classification and use a schema-like solution (as the person solving the motion problem above did). Although the number of observations in this protocol study is small, the greater number of schema-like solutions in the appropriate condition replicates Hinsley et al.'s (1977) finding.

In Experiment 1a, two main results occurred in problem-solving performance and were supplemented by the protocol data in Experiment 1b. First, the inappropriate condition led to lower performance than the other two conditions, indicating an effect of content. At least some of this poor performance may have been due to a miscategorization of the problem on the basis of the content. Second, the appropriate and neutral conditions did not differ in accuracy or latency. The protocols, however, did suggest that the appropriate condition may lead to earlier access of the relevant knowledge of the problem type. To examine this possibility, it is useful to have further data on problem categorization.

The plan for the rest of this article is to further analyze these content effects. In Experiment 2, we examined problem categorization, which is often claimed to be a crucial component of problem solving (e.g., Chi et al., 1981). We used an on-line measure to find out what categories the solvers were thinking of at different points in reading the problem. We hoped to gain a better understanding of whether the inappropriate content initially misleads the solvers and of whether there is an appropriate–neutral difference here. In Experiments 3a and 3b, we aimed for a better understanding of the appropriate and neutral conditions' similarities and differences, by focusing on how the relevance of problem information is determined.

Experiment 2

A critical part of problem solving is how solvers determine the problem category, but the effects of problem content on categorization, or schema access, are not well understood. As discussed in the introduction, earlier results have shown that novices are likely to use content and experienced solvers are likely to use the deep structure of problems when sorting problems. Chi et al. (1981) found that when faced with problems varying in content and deep structure, experienced solvers grouped the problems by deep structure while novices grouped by content (see also Silver, 1979). However, these results occurred when content and structure were varied orthogonally and involved a sorting task in which the problems could be explicitly compared with one another with no time pressure. By orthogonally varying content and deep structure, any influence of content on categorization could be seen only if it was greater than the influence of deep structure. The task itself, although providing some insights into how solvers view problem similarity, did not directly address how the content may be used while actually reading the problem. This issue of how problems are categorized on-line is critical to understand, because once a problem has been categorized, the solution procedure used to solve it may already have been determined.

Hinsley et al. (1977) examined on-line categorization by presenting a phrase at a time and asking people to categorize the problem after each phrase had been presented. They found that when the content was appropriate for the problem type, people could quickly identify the problem category, often before any words relating to the deep structure of the problem had been presented. In addition to the category to which the problem belonged, people were able to provide information relevant to the solution of such a problem. This finding suggests that content may be used by experienced solvers to access knowledge about the problem type, including knowledge about the solution procedure. Hinsley et al. used only problems that had what we are calling appropriate contents, so their results do not provide information about how variations in content may affect on-line problem categorizations. The purpose of the current experiment was to explore the effects of content in an on-line categorization task by varying the relation of content to problem type. We are not aware of other research that has examined this issue.

How does the content affect the amount of the problem needed to categorize it? If solvers are using content as a quick heuristic cue to the problem type, then the appropriate condition problems will be categorized faster than the neutral, which will in turn be categorized faster than the inappropriate. However, the neutral condition problem-solving performance was as high as in the appropriate condition in Experiment 1, so the categorization may not be different either.

In addition to determining if a difference existed among the three content conditions, we had two further questions about the influence of content on categorization. First, are problems categorized correctly in all conditions by the final phrase? Perhaps experienced participants are temporarily affected by content, but can recover with additional problem information. This result might help explain the lack of the appropriate-neutral difference in Experiment 1. Second, what is the pattern of categorizations in the inappropriate condition in which problem type and content point to different schemata? As proposed earlier, solvers may be misled by the inappropriate content. If so, we should be able to detect these misled categorizations here. The main result and the answers to these two questions provide important information about the role of content in problem categorization.

Method

Participants. The participants were 12 graduates of IMSA who were attending the University of Illinois and had not participated in Experiment 1. They were paid $5 for their participation in the experiment, which lasted about 1 hr.

Materials. The problems were the same ones used in Experiment 1. For this experiment, however, each problem was broken into between five and nine phrases, with most problems split into seven phrases. The last phrase of the problem was always the question that the problem posed. Table 3 contains the phrases for the work problem type for all three conditions. The matched problems across the three conditions were split into the same number of phrases, with the corresponding phrases of the problems containing the same information. Each phrase was printed on a 1 in. × 8.5 in. slip of paper.

Procedure. Participants were tested individually and their answers were tape-recorded for later transcription. Participants were given 15 problems, 3 practice and 12 test problems. The method of distributing problems across participants was the same in this experiment as in Experiment 1a. Every problem was presented to 4 participants.

The trial for each problem began with the participant receiving the first phrase. The participant read the phrase aloud and was then asked three questions: (a) How would you categorize this problem? (b) What sort of information do you expect in later phrases? and (c) What will the final question be? After responding to the questions, the participants were given the second phrase of the problem, and the same questions were asked. The experimenter gave no feedback to the participants concerning the correctness of their responses, but continued to give the participants the phrases of the problem, one at a time and in order, until either no phrases remained or until the participant clearly knew what the problem entailed (i.e., giving complete answers to all three probe questions). In answering Question b, participants would often give several possibilities as to what may come in later phrases, and they would often also give a couple of answers to Question c. For example, if an interest problem started out "An investor . . . ," participants who recognized the problem type right away would perhaps say, "This is one of those interest problems, where the person puts money in one account at such and such an interest rate, and probably some more money in another account for some amount of time, and you have to find either the total amount of money or the interest rates." Participants were allowed to keep the phrases they had already received from a problem in view. After participants completed all 15 problems, the experimenter asked them questions to clarify any unclear or ambiguous statements they had made when answering the three questions.

Table 3
Sample Work Problem Split Into Clauses for Experiment 2

Appropriate content
Two workers, Jane and Abby, . . .
. . . work in a crayon factory. . . .
. . . Jane can fill a 24-count box of crayons . . .
. . . in 5 min. . . .
. . . Abby can fill a 24-count box . . .
. . . in 8 min. . . .
. . . Their boss decides to have them work together . . .
. . . filling 36-count boxes. . . .
. . . How long will it take them to fill a 36-count box?
Neutral content
Two fishermen, Jim and Tom, . . .
. . . enter the Fairfield fishing contest. . . .
. . . Jim can catch 24 fish . . .
. . . in 5 hrs. . . .
. . . Tom can catch 24 fish . . .
. . . in 8 hrs. . . .
. . . This year they enter the two-man division . . .
. . . in which they much catch 36 fish together. . . .
. . . How long will it take them to catch 36 fish?
Inappropriate content
A riverboat . . .
. . . has 2 engines. . . .
. . . Its main engine running alone can propel the boat 24 miles downstream . . .
. . . in 5 hours. . . .
. . . The backup engine by itself can take the boat 24 miles downstream . . .
. . . in 8 hours. . . .
. . . One day the boat goes with both engines running . . .
. . . 36 miles downstream. . . .
. . . How long does this trip take?

Results

The problems were scored according to the proportion of the problem's phrases the participants needed before they correctly categorized the problem. A sufficient answer contained enough information to completely answer the second probe question, and either the first or third probe question.[3] Examples of answers to the first probe question would be "This is a riverboat problem" or "This is one of those age problems." For the second probe question, we expected explanations similar to those used by Mayer (1981) in his example templates (e.g., "One vehicle starts and is followed later by a second vehicle that travels over the same route at a faster rate" for an overtake motion problem). It was usually the case, however, that when participants did correctly categorize a problem, the information they provided answered all three questions. Because the names of problem categories are somewhat idiosyncratic, the determination of whether a category was correct sometimes had to depend on the comments the participants made after testing. For example, if a participant classified an interest problem as a "bank problem," and after the experiment explained that bank problems meant problems dealing with interest rates and investors depositing an amount of money in different accounts, that information would be considered in scoring that problem. We adopted this procedure to avoid giving the appropriate condition an unfair advantage of being able to simply use the content as a category whether they knew the problem type or not. Because many algebra word problem types are referred to by their content (e.g., age, interest, work), it may not be possible to totally eliminate this appropriate advantage, but this procedure did require a clear explanation of the type.

We both scored the answers, and we again had a person blind to condition score the answers. The independent scorer was slightly more conservative, but 82% of her rankings were either identical to or different from our judgments by one phrase. Furthermore, an analysis based on her scoring provided results similar to those presented below.[4]

[3] We thank an anonymous reviewer for suggesting this scoring method.

[4] The means from her scoring were .30, .60, .78 for the appropriate, neutral, and inappropriate content problems, respectively, $F(2, 11) = 32.08, MSE = 0.157, p < .01$.

The main result of interest is the proportion of phrases seen before correctly categorizing a problem in the three conditions. Participants required a mean of .29 of an appropriate content problem in order to correctly categorize it, .55 of a neutral content problem, and .79 of an inappropriate content problem. A repeated-measures analysis of variance showed that this is a significant difference, $F(2, 11) = 55.42$, $MSE = 0.208$, $p < .01$, and a Newman-Keuls test indicated that all means differ from one another ($p < .01$).[5] This exact pattern (appropriate < neutral < inappropriate) occurred for 11 of the 12 participants, and for 10 of the 12 base problems.

In addition to this main difference of conditions, the two other questions may now be answered. First, were problems correctly categorized by the final phrase, when the whole problem had been read? This measure too shows some effect of content. Participants always (48 of 48) categorized the appropriate content problems before the final phrase and almost always had the correct category before the end with the neutral content problems (44 out of 48). Performance was lower in the inappropriate content problems (35 of 48), though they still were correctly categorized almost 75% of the time. A sign-test by participants showed the difference between the neutral and inappropriate problems to be marginal ($p = .07$).

Second, what is the pattern of categorization with the inappropriate content problems? Because the content was available in the first phrase or two, participants initially would often (37 out of 48 times) categorize the problem incorrectly on the basis of the inappropriate content. In many cases, they then realized that there was a difficulty, that the problem did not seem to be of the type they had initially thought. Often, they were able to recover and make an appropriate categorization, but in some cases they remained confused to the end. Incorrect categorizations never occurred with the appropriate or neutral content problems.

Discussion

This experiment demonstrated that solvers could use a problem's content to correctly categorize the problem. In contrast to Experiment 1, performance differed among all three content conditions. The extent to which a problem's content matched the problem's underlying type affected the proportion of the problem statement the participant needed to read before making a correct categorization. If a problem's content was consistent with the problem's type, then fewer phrases of the problem were required than if the problem's content reflected no particular problem type. However, most participants were able to categorize the problems before the last phrase, the question, of the problem was read, even in almost 75% of the inappropriate content problems. At least some category information was available to the participants after having read the problem once through.

This could be the reason why the appropriate and neutral conditions did not differ in Experiment 1a. Even though participants were usually faster to categorize the appropriate content problems, the correct category information had been deduced by the time they had read all of an appropriate or neutral content problem. Therefore, at the outset of actually trying to solve the problem, the information available to them was the same. In this light, it is not surprising that we found no differences between the appropriate and neutral content problems in Experiment 1a. It was only with the inappropriate content problems that participants did not have category information available to them after having read the problem, and so their performance suffered when trying to solve the problem.

As a last note, although this task was very different from how people usually read a problem, we think it provides a reasonable experimental technique for gauging what knowledge is available to the solvers during reading (see Rumelhart, 1981, for a related technique for examining reading). First, it was not viewed as strange by the participants. They were all clear about the idea of problem types, and seemed comfortable with trying to classify these problems into types. Second, in protocols, solvers often mention problem types during reading of the problem. Apparently students are either taught or develop these categories when they learn algebra and have information stored about these categories that enables them to identify problem types by seeing only part of the problem.

Experiments 3a and 3b

In the two parts of this final experiment, we further analyzed content effects by examining another crucial part of problem solving that is also thought to be an important use of problem schemata—the determination of relevance of problem information. If problems contain much information that is not necessary for the solution, what allows the solvers to figure out what is relevant and what is not? Deciding about the relevance of information in problems is often thought to be a common part of understanding the problem and preparing to solve it (e.g., Hayes, Waterman, & Robinson, 1977).

The access of a schema, or other relevant problem-solving knowledge, might allow a determination of what information is likely to be relevant. As shown in Experiment 2, classification of problems into categories could occur without the entire problem being read, so an early categorization may influence how the information presented in the problem is used. In addition, the variables within a schema might be content sensitive. That is, once a particular schema has been activated by a problem, the variables and slots in the schema may be receptive only to certain types of values. For example, once a problem has been identified as a river current problem, the variable representing the rate of the current may be bound to a quantity named to indicate that it is, in fact, the rate of the current rather than some other rate (such as the rate of an escalator, as in our neutral content river current problems).

---

[5] The results were very similar when only the first half of each participant's data was examined: They required a mean of .31 for an appropriate content problem, .60 for a neutral content problem, and .84 for an inappropriate content problem, $F(2, 11) = 36.30$, $MSE = 0.093$, $p < .01$.

Once again, the three problem types from three different families showed the same pattern of performance as the three types from the same (time rate) family. For the different families, performance was .35, .57, and .77 for the appropriate, neutral, and inappropriate conditions, respectively. The corresponding numbers for the same family problem types were .23, .53, and .80.

Thus, having a problem schema might help the problem solvers bind numbers to variables by helping them to reject irrelevant variables.

Hinsley et al. (1977) discussed a similar experiment, in which they had 6 participants solve the "Smalltown Problem." The Smalltown Problem was a complicated algebra motion problem that contained irrelevant information. Part of the irrelevant information could be construed as coming from a "triangle" algebra problem. Half of their participants attended to the irrelevant information and attempted to use that information in formulating the problem (while apparently still in the act of reading the problem). In the discussion of their results, Hinsley et al. mentioned that some participants even misread parts of the problems (e.g., *minutes* for *miles*). These findings bolster the claim that participants do categorize problems while reading them, and that they use these on-line categorizations to help solve the problems. However, given the small number of observations (six) on a single problem, the impact of irrelevant information is still not clear.

Our focus in this experiment was to use a similar design to better understand the appropriate and neutral condition differences. Experiment 2 showed that appropriate problems are categorized much earlier, but as seen in Experiment 1a problem solving performance was about the same for the two conditions. After the first experiments, we mentioned a number of possible reasons for this lack of difference. In Experiment 3, we look at one possibility in more detail—problem complexity.

More complex problems (i.e., those with irrelevant information) might lead to an appropriate–neutral difference, because early access of the relevant knowledge or schema might be of great help in determining relevance of the problem parts as the problems are being read and initially instantiated. In Experiment 3a, we focused on the appropriate and neutral conditions (where simple problems are almost always categorized by the end of reading) and examined the effects with more complex problems. We did not use the inappropriate content problems because we wanted to concentrate on the lack of difference found in Experiment 1a, and by removing the inappropriate content condition, each participant could receive a higher proportion of appropriate and neutral content problems. To this end, the appropriate and neutral problems used in Experiments 1 and 2 were made more complex by adding irrelevant information. The added information fit into the story line of the problem, and so could not easily be discarded, but did not affect the problem's solution. Unlike Hinsley et al.'s (1977) Smalltown Problem, the irrelevant information did not come from another problem type, but rather from the same problem type. If content is of use in solving problems, then the irrelevant information in the appropriate content problems should be passed over more easily than in the neutral content problems. This difference in sensitivity to relevance should result in more accurate and faster solutions in the appropriate content problems.

It is important to note that by "complex," we are referring to the addition of irrelevant information, not the complexity of the deep structure. Our goal in this addition of complexity was not to force the problem solvers to adapt their problem types, but rather to force a more extended use of these schemata so that any different effects of content may be more readily observed.

Experiment 3a

Method

Participants. The participants were 16 graduates of IMSA attending the University of Illinois who had not participated in either Experiment 1 or 2. They were paid $5 for their participation in the experiment, which lasted about 45 min.

Materials. The problems used were modified versions of the appropriate and neutral content problems from Experiments 1 and 2. Inappropriate content problems were not used. The added information was not needed to solve the problem, but was information that could not simply be discarded because it seemed out of place. Whenever possible, the extra information contained numbers with associated units similar to the ones used to actually solve the problem and similar sorts of irrelevant information were added to both the appropriate problem and its matching neutral content problem. Appendix B illustrates how the problems from Appendix A were changed for this experiment.

Procedure. Participants were tested in small groups of 4–6. The testing procedure was similar to that of Experiment 1a. Participants were given booklets of 15 problems, 1 problem per page. The first 3 problems were the practice problems, which also had irrelevant information added to them. Then there were the 12 test problems, the modified appropriate and neutral content problems, 1 from each of the base problems. In this counterbalancing, every problem was presented to 8 participants. As in Experiment 1a, after every 45 s, the experimenter would call out, "Line," and the participants would draw a line across their page and continue work below that line. Participants had a maximum of 3 min to work on every problem and were not permitted to work ahead.

Results

The scoring was as in Experiment 1a. After conducting the experiment, we found that there were serious wording difficulties with two problems (one appropriate interest and one neutral motion problem), making the problems ambiguous and exceptionally difficult (or impossible) to solve. The presented scores do not include these problems (and their matched problems), but the results were very similar when they were included and are given in footnotes.

These complex materials did show an advantage for the appropriate condition. Participants scored a mean of .66 on the appropriate content problems and .58 on the neutral content problems, $t(15) = 2.13$, $SEM = 0.038$, $p = .05$. This advantage of appropriate content was found for 13 of 16 participants and for 7 of 10 base problems. The time measure showed only a small, nonsignificant advantage for the appropriate content problems, with a mean of 1.76 min versus 1.90 min for the neutral content problems, $t(15) = -1.60$, $SEM = 0.088$, $p > .1$, and was found for 8 of 16 participants and for 7 of 10 base problems.[6]

[6] Including the two excluded problems, the accuracy measures for the appropriate and neutral content problems were .66 and .57, respectively, $t(15) = 2.64$, $SEM = 0.034$, $p < .05$, and the time measure was 1.87 min for the appropriate content problems and 1.99 min for the neutral content problems, $t(15) = -1.01$, $SEM = 0.129$, $p > .1$. In

In performing an analysis similar to that in Experiment 1a, in which we examined the latency only for the problems for which participants scored .75 or higher, we found a similar pattern. There were 59 appropriate content problems that met that criterion, and participants spent 1.49 min solving them, compared with 48 neutral problems, which participants spent 1.59 min solving. Again, participants were slightly faster at solving appropriate content problems, even though there were fewer neutral problems that met this criterion.

Discussion

With the added irrelevant information, a difference was detected between the appropriate and the neutral content problems, in contrast to the simpler problems used in Experiment 1. Although the problem solvers were experienced in solving algebra word problems, the appropriate content problems were solved more accurately. Content does affect how people determine the relevance of problem information.

Experiment 3b

As with Experiment 1a, a follow-up experiment was performed in which participants were asked to talk aloud while solving the problems. An analysis of these protocols may shed more light on how a problem's content affects the application of a schema. The question of most interest is whether the content of the problem changes the solvers' sensitivity to the relevance of the material. The advantage of using these complex problems is that we can examine the protocol for how often the relevant versus irrelevant aspects of the problems were talked about as a function of content. If content is having an effect on instantiating the schema, then the appropriate problems should include more discussion of the relevant aspects and less discussion of the irrelevant aspects than the neutral problems.

Method

Participants. The participants were 8 graduates of IMSA attending the University of Illinois who had not participated in any of the previous experiments. They were paid $5 for their participation in the experiment, which lasted a little over an hour.

Materials. The problems were the same complex problems used in Experiment 3a. The wording difficulties associated with two of the problems from that experiment were fixed.

Procedure. The participants were tested individually and their comments tape-recorded and transcribed later. Prior to the experiment, the participants were instructed on how to give a talk-aloud protocol and were given a practice talk-aloud task. Participants were given a booklet with 13 problems: 1 practice problem and 12 test problems. Every problem had four observations. Participants had paper and pencil to assist them in solving the problem. If participants were quiet for a period of time, the experimenter reminded them to keep talking.

Results and Discussion

For the objective performance measures, the participants scored .67 on the appropriate content problems and .65 on the neutral content problems, $t(7) = 0.34$, $SEM = 0.059$, $p > .1$. Although the accuracy result is nonsignificant, 6 of the 8 participants did perform better on the appropriate content problems (with an average difference of .08). The 2 remaining participants performed much worse on the appropriate problems (with differences of .17 and .21), leading to little overall difference. These 2 participants performed extremely well overall, but each had trouble with two of the appropriate problems. For the latency measure, there was a slight difference, with participants taking 2.95 min on average to solve the appropriate content problems, and 3.68 min to solve the neutral content problems, $t(7) = -1.97$, $SEM = 0.371$, $p < .1$. Although the results do not exactly match those from the main experiment, there does still seem to be the main result of an advantage for the appropriate condition over the neutral condition.

As in Experiment 1, we rated each solution as either schema-like or sentence-to-equation. Also as in Experiment 1, participants were more likely to use a schema for the appropriate content problems (30 out of 48 cases) than for the neutral content problems (20 out of 48 cases), $t(7) = 2.35$, $SEM = 0.089$, $p = .05$. We again had a scorer blind to condition rate the solution methods, and her ratings coincided with Stephen B. Blessing's 90 out of 96 times (.94).

The main purpose of collecting these protocols was to examine how much time the participants concentrated on the relevant aspects of the problem versus the irrelevant parts of the problem in formulating their solution. To this end, the protocols were divided into lines. A line was either a complete sentence that the participant spoke concerning one topic, or one incomplete sentence bounded by a 2-s pause. In the rare instance when one line contained a combination of relevant, irrelevant, or miscellaneous information (discussed shortly), the sentence was divided accordingly. Every line was coded as either mentioning relevant information, irrelevant information, or just a miscellaneous comment. A relevant line would be "⅓ of the total rabbits, T over 3 we'll call it," or "distance equals rate times time." Irrelevant lines resemble the relevant lines, but contain numbers referring to the irrelevant information. Miscellaneous lines include all of the "okays" and "let's sees," and comments about the problem, such as "Oh, that information is irrelevant," or "I'm not going to be confused by all this junk you're putting in these." These statements could be classified as relevant, because by stating such, participants are indicating they are not regarding that irrelevant informa-

---

analyzing only the first six problems that each participant received (one of each type), we found the results to be similar: Participants scored an average of .64 on the appropriate content problems and .53 on the neutral content problems, $t(15) = 3.73$, $SEM = 0.029$, $p < .05$, and they solved the appropriate problems in 1.82 min and the neutral problems in 2.12 min, $t(15) = -1.45$, $SEM = 0.207$, $p > .1$.

The three problem types from three different families again showed the same pattern of performance as the three types from the same (time rate) family. For the different families, accuracy was .67 and .62, with latencies of 1.80 and 1.92. The corresponding numbers for the same family problem types were, for accuracy, .64 and .55, and for latency, 1.72 and 1.88.

tion further, which is the proper response. However, to be strict we put such statements in the miscellaneous category because they did not specifically mention relevant aspects of the problem. Also, these results, like the time measure from Experiments 1a and 3a, are only for the lines between when the participant read the problem's question to when the participant wrote the equation, correct or incorrect, that they used to solve the problem. The interrater reliability between Stephen B. Blessing's ratings and the independent scorer was 95%. Of most interest is whether the difference in the proportion of lines discussing relevant versus irrelevant information was greater for the appropriate condition than the neutral condition.

The top section of Table 4 displays the results of this analysis. Participants were more sensitive to the relevance of the problem information in the appropriate condition than in the neutral condition. Relevant information was discussed marginally more often in the appropriate condition than in the neutral condition, $t(7) = 2.25$, $SEM = 0.031$, $p < .06$, and irrelevant information was discussed marginally less often, $t(7) = 2.30$, $SEM = 0.026$, $p < .06$. Combining these to get an overall measure of sensitivity to relevant versus irrelevant information, the difference of .48 in the appropriate condition was significantly greater than the difference of .35 in the neutral condition, $t(7) = 2.40$, $SEM = 0.054$, $p < .05$. This difference was positive for all 8 participants. Thus, the appropriate content condition led to greater sensitivity to the relevance of the problem information.

These relevance judgments can be looked at more closely, though as we examine subparts of the data, the number of observations is smaller. If schemata are allowing people to more readily see the relevance of information, then the effects of relevance sensitivity should be greater when participants actually used a schema in solving the problem. That is, for those problems when participants used the sentence-to-equation method, there should be less difference between the relevant and irrelevant information measures than for those problems when they used the schemata. There is only a small nonsignificant difference in this predicted direction, with a mean relevance minus irrelevance measure of .46 for the problems when schemata appeared to be used versus .39 for problems when participants appeared to use the sentence-to-equation method, $F(1, 20) < 1$, $MSE = 0.040$.[7] We feel that these results are too preliminary to draw strong conclusions from, but the means are at least consistent with the predicted pattern. The data are further broken down by the appropriate and neutral conditions in the bottom two panels of Table 4. Interestingly, as can be seen in the bottom panels of Table 4, across both the schema and sentence-to-equation cases, the appropriate condition problems led to a marginally greater relevance effect, $F(1, 20) = 3.24$, $MSE = 0.040$, $p < .10$. The difference in the schema cases was twice as large as in the sentence-to-equation cases (0.18 vs. 0.09), but the interaction was not near statistical significance, $F(1, 20) < 1$. The appropriate advantage in the schema cases suggests the possibility that the schemata may include surface content information, so that finding relevant information in the appropriate content problems is facilitated. This advantage in the sentence-to-equation cases is more difficult to interpret, but it may reflect some times in which solvers had a schema but used a sentence-to-equation method for solving the problem.

The protocols collected during this experiment can also be used to make the point that categorizing a problem can aid in the determination of problem statement relevance. In questioning the participants after the experiment, all except for one reported that they tried to categorize the problems into types, and that this categorization helped. For example, one participant while solving a motion problem said, "So I'm thinking rate times time equals distance again." After that, she immediately picked out the three necessary pieces of information from the problem statement and proceeded to set up the correct equation. Another interesting case of this occurred when a participant was solving a neutral interest problem, where the content was about flowers: "And if she's got—this is like an interest problem. She's got 190 new plants with flowers total so the number of red flowers plus the number of blue flowers is equal to 2,000 plants, and then you've got 190 new plants which equals 8% of the red plus 10% of the blue." Once she figured out this was an interest problem, she was able to quickly set up the equations needed to solve the problem. In her comments after the experiment, she wrote, "You always learn them as interest problems in math class so you always call them that, even if you use them differently. You don't group flowers as money but both problems are solved the same way."

Many participants adopted the strategy of reading the question first, once they figured out that the problems contained irrelevant information. They would then read through the rest of the problem, determining if the current statement could be used in solving the problem. Many participants (5 out

Table 4
Proportion of Lines in Each Statement Category for the Complex Problem Protocols of Experienced Participants (Experiment 3b)

Analysis type	Miscellaneous	Relevant	Irrelevant	Relevant − Irrelevant[a]
Combined[b]				
Appropriate	.17	.66	.17	.48
Neutral	.18	.59	.23	.35
Separated by solution method				
Schema				
Appropriate (30 cases)	.15	.70	.15	.55
Neutral (20 cases)	.17	.60	.23	.37
Sentence-to-equation				
Appropriate (18 cases)	.21	.61	.18	.43
Neutral (28 cases)	.18	.58	.24	.34

[a]Relevant − irrelevant is a measure of the sensitivity to the relevance of the information. [b]Combined includes the problems solved by both the schema and sentence-to-equation. There are 48 problems in both conditions.

[7] We thank an anonymous reviewer for suggesting this analysis.

of 8) physically crossed out the information they thought to be irrelevant. Participants would sometimes categorize the problems while still reading the problem statement and in some cases would begin to formulate the problem (e.g., "Wendy is currently 4 times older than her niece so $n = x$ and $w = 4x$.").

The results of Experiments 3a and 3b show two effects of appropriate versus neutral content with these more complex contents. First, in Experiment 3a, performance was higher with appropriate contents. Second, in Experiment 3b, participants were more sensitive to the relevance of information with appropriate content. Taken together, these results suggest that with more complex problems, the appropriate content allows solvers to more easily focus on the relevant problem aspects and solve the problem.

General Discussion

These studies show a clear effect of content on the problem solving and problem categorization of experienced problem solvers. Most of the previous work examining content effects has focused on novices. These results suggest that even experienced problem solvers have knowledge about problem types that is sensitive to formally irrelevant, but empirically predictive, contents. The contents, because they are available so quickly, may allow such experienced solvers to activate the relevant knowledge about the problem type to help in understanding and solving the problem.

Much of the work on experienced problem solvers suggests that experience in the domain has led these solvers to build up problem schemata. We will now consider what these results might suggest about two important issues in this literature: the access versus application of schemata and the acquisition of schemata.

Schema Access Versus Application

Content effects in problem solving might be due to either the access or application of problem schemata, or both. The problem's content, because it is readily apparent, might allow the experienced problem solver to categorize it (i.e., access the appropriate schema) early in reading the problem. Quick access on the basis of surface content, even if it is not guaranteed to be correct, may be an attractive initial hypothesis given the longer time required for determining the deep structure. The importance of this quick access is that experts would be able to begin formulating the problem for solution while still reading it, thus saving time in solving the problem. It would be a strange expert who could not take advantage of the strong predictive relationship between content and deep structure (Lewis & Anderson, 1985).

Work done by Novick (1988) provides evidence that experts may indeed initially base their initial categorizations on a problem's surface content. In many cases, the experts in her study initially tried a solution method that was successful on a previous problem with a similar surface content. Only after failure with the previous solution method on the current problem did the experts attempt a different solution method. However, when one of the previous problems also shared a similar deep structure to the current problem, experts were more likely than novices to use the solution method indicated by that deep structure.

As far as we can tell, all of the main results in our experiments can be accounted for by content effects on schema access. In Experiment 2, such access differences seem to be the likely explanation for the fact that the proportion of the problem needed to categorize a problem varied with the appropriateness of the content. In Experiment 1a, we found that the inappropriate problems were solved less accurately and more slowly than the other two types of problems, which did not differ from one another. On the basis of the final categorizations of Experiment 2, it is possible that this difference was solely due to access differences; by the final phrase almost all the appropriate and neutral problems were correctly categorized, while about 25% of the inappropriate problems were not. Experiment 1b is more problematic for this access explanation, because we found a difference in solution methods for appropriate and neutral problems, though the number of observations making up this difference was small. In addition, it is possible that early access (as was more probable in the appropriate condition) is needed to use a schema-like solution (i.e., the schema needs to be accessed early enough to avoid beginning a sentence-to-equation translation). Experiment 3a, with complex problems, showed clear differences between the appropriate and neutral problems, but again these could be accounted for by access differences. With complex problems, earlier access by the problem content may be especially helpful in allowing the solver to determine the relevance of the various pieces of information given in the problem. This earlier access might then lead to better performance (Experiment 3a) and a greater sensitivity to the relevance of problem information (Experiment 3b). Thus, the main results of these studies do not require an explanation beyond the effects of content on schema access.

Although our results can be explained in terms of schema access, the effects of content may be more pervasive, affecting even the application of the schema. The content may allow easier instantiation of the schema if the schema includes some content specific information (e.g., the rate of the river current for a river current problem). If that were true, it would mean that the content was not simply a retrieval cue or trigger for the schema, but was in some way embedded within the schema. For example, with a river current schema, perhaps the variables for the schema are not simply the object's rate and the rate of a helping or hindering force, but instead may be the boat's rate and the river current rate. Such content embedding amounts to a specialization of the schema, so that it would be easier to apply to typical river current problems (because the mapping of variables could be done at the surface level), although it would be more difficult to map problems without the typical contents. Research by Bassok and Holyoak (1989; Bassok, 1990) examining transfer between algebra and physics is extremely suggestive that this is the case. Such a trade-off may be a good one, depending on the relative costs and benefits, as well as the probabilities of typical and atypical contents (e.g., Shavlik, DeJong, & Ross, 1987, present similar arguments for intermediate generalizations in physics prob-

lems. Also, Allen & Brooks, 1991, and Rothkopf & Dashen, 1995, present some related ideas on specializations of categorization rules).

Schema Acquisition

The current modal view of expertise is that experts are able to solve problems expertly because they have many domain-dependent, highly-specific problem schemata. From this perspective, a crucial issue becomes how such schemata are acquired (e.g., Reimann & Chi, 1989). Our reading of the current work on learning suggests that there are likely to be multiple ways in which schemata are acquired, but we think that content effects may be helpful in thinking about acquisition. As an example, we consider one of the simpler learning suggestions, that problem schemata are built up by comparing solved problems of a given type and extracting their commonalties. (Such a view has been espoused in a number of different ways by Anderson, Kline, & Beasley, 1979; Carbonell, 1983; Cummins, 1992.) An important distinction among these views is how people know which problems to compare. In some views, this information is given by helpful teachers or programmers (e.g., Gick & Holyoak, 1980; Gick & Holyoak, 1983; Rumelhart & Norman, 1981), but in many real-world cases the learner needs to determine which problems to compare.

Content, and especially the correlation between content and problem type, provides one possible answer (Ross & Kennedy, 1990; see Bassok, 1990, and Novick & Holyoak, 1991, for related ideas). Examining the results of this study, it is extremely probable that the problem schemata possessed by our experienced solvers contained content cues. We know from much work on analogical problem solving that surface problem features (i.e., content) are potent reminders of earlier problems with similar contents (e.g., Gentner & Landers, 1985; Holyoak & Koh, 1987; Ross, 1984). If this surface feature similarity led to remembering the earlier problem, it would allow comparisons among problems with similar contents. If problems with similar contents occurred often for a given problem type and later problems helped retrieve generalizations learned from earlier comparisons, it might lead to problem schemata that included contents (at least as a retrieval cue but, depending upon its predictiveness, perhaps even embedded within the schema). Bernardo (1994) provides evidence for this view. He found that people are conservative when they make generalizations—information concerning content is not abstracted away. An alternative to the generalization-from-use idea would be to compare solved problems (e.g., Cummins, 1992), but again using content as an influence on which problems to compare. In either case, early reliance on content may be quite useful when a learner is struggling to learn the concepts, categories, and procedures used in a particular domain, though clearly its usefulness depends upon the predictiveness of the content.

Conclusions

From these results, it is clear that content does play an important role in solving and categorizing algebra word problems, even for experienced problem solvers. When a problem's content is inconsistent with the problem's underlying type, then experienced solvers are less accurate and slower at solving the problem and may even not come up with a correct categorization. When a problem's content is consistent with the underlying type, experienced solvers are faster to categorize the problem and are more sensitive to the relevance of problem information for complex problems. We interpreted these results largely in terms of how content may provide access to appropriate schematic knowledge, though it is also possible that some of the results may be due to effects on schema application.

References

Allen, S. W., & Brooks, L. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General, 120,* 3–19.

Anderson, J. R., Kline, P. G., & Beasley, C. M. (1979). A general learning theory and its applications to schema abstraction. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 13, pp. 227–318). New York: Academic Press.

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

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.

Bernardo, A. B. I. (1994). Problem-specific information and the development of problem-type schemata. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 379–395.

Carbonell, J. G. (1983). Learning by analogy: Formulating and generalizing plans from past experience. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 1, pp. 137–161). Palo Alto, CA: Tioga.

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.

Cummins, D. D. (1992). The role of analogical reasoning in the induction of problem categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1103–1124.

Gentner, D., & Landers, R. (1985). Analogical reminding: A good match is hard to find. In *Proceedings of the International Conference on Systems, Man, and Cybernetics* (pp. 607–613). New York: Institute of Electrical and Electronics Engineers.

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.

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.

Hayes, J. R., Waterman, D. A., & Robinson, C. S. (1977). Identifying the relevant aspects of a problem text. *Cognitive Science, 1,* 297–313.

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. A. 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.

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. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. *Instructional Science, 10,* 135–175.

Mayer, R. E. (1982). Memory for algebra story problems. *Journal of Educational Psychology, 74,* 199–216.

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.

Reimann, P., & Chi, M. T. H. (1989). Human expertise. In K. J. Gilhooly (Ed.), *Human and machine problem solving* (pp. 161–191). New York: Plenum Press.

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

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.

Rothkopf, E. Z., & Dashen, M. L. (1995). Particularization: Inductive speeding of rule-governed decisions by narrow application experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 469–482.

Rumelhart, D. E. (1981). *Understanding understanding* (Tech. Rep. No. 100). University of California, San Diego, Center for Human Information Processing.

Rumelhart, D. E., & Norman, D. A. (1981). Analogical processes in learning. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 335–360). Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99–135). Hillsdale, NJ: Erlbaum.

Schoenfeld, A. Y., & Hermann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 484–494.

Shavlik, J. W., DeJong, G. F., & Ross, B. H. (1987). Acquiring special case schemata in explanation-based learning. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 851–860). Hillsdale, NJ: Erlbaum.

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

Appendix A

Examples of Materials Used in Experiments 1 and 2

Age Problem

Appropriate

Michelle is 4 times older than her niece. In 5 years, Michelle will be 3 times older than her niece. How many years older is Michelle than her niece?

Neutral

Wendy got 4 times as many points on an exam than Dan got. The teacher gave everyone 5 more points. After that Wendy had 3 times as many points as Dan. How many more points than Dan did Wendy score?

Inappropriate

A mason mixed 4 times as much cement in one container as another. He adds 5 liters of cement to each mixer. Now the first has 3 times the cement as the second. How much does each contain now?

Mixture Problem

Appropriate

A chemist mixes two types of solutions. One solution contains 20% boric acid. The other solution contains 30% lactic acid. His new solution fills a 10 pint jar and is 23% acid. How much of each of the original solutions did the chemist pour in the jar?

Neutral

Bob, a partygoer, went to a big party last night and drank two types of punch. One punch was made with 20% pineapple juice. The other punch was made with 30% orange juice. By the end of the evening, Bob drank 10 pints of punch, 23% of which was fruit juice. How much of each type of punch did Bob drink?

Inappropriate

Bart went to several birthday parties. Some friends were turning 20, and the rest turned 30. Bart went to 10 parties, and the average age of the birthday person was 23. How many of Bart's friends turned 20?

Motion Problem

Appropriate

Two drivers leave for Los Angeles at the same time. George starts out 72 miles from LA and Peggy starts out 100 miles from LA. Both reach LA at exactly the same time. George drives at a speed of 27 mph. How fast does Peggy drive?

Neutral

Two archers fire their arrows at the same target at the same time. Phil is standing 72 meters from the target and Rudy is standing 100 meters from the target. Both arrows hit the target at exactly the same time. Phil's arrow flies at a speed of 27 meters/sec. How fast does Rudy's arrow fly?

Inappropriate

Two investors, George and Peggy, hold different stocks. George needs $72 more to make his first thousand, while Peggy needs $100. They reach their first thousand at the same time. George made $27 each day. How much did Peggy make each day?

Work Problem

Appropriate

An electrician can complete a job in 2 hrs. His apprentice takes 4 hrs to complete the same job. The electrician and his apprentice work on the job together. How long will it take them to do the job?

Neutral

A pair of trout can fill a pond with their offspring in 2 months. A pair of carp take 4 months to fill the same pond. A pair of trout and a pair of carp are put into the pond together. How long will it take them to fill the pond?

Inappropriate

A tugboat pushes a ship 10 miles upstream in 2 hours. Another tugboat could push the ship 10 miles upstream in 4 hours. If they worked together, how long would that task take?

Appendix B

Examples of Complex Materials Used in Experiment 3

Age Problem

Appropriate

Wendy is currently 4 times older than her niece. Wendy is also 6 years older than her youngest sister, Rachel. Rachel wants to go to Disney World before she is 15 years old. Rachel's friend went to Disney World when she was exactly twice as old as Rachel is now. In 5 years, Wendy will be 3 times older than her niece. Rachel hasn't gone to Disney World. The average age of Rachel's family is 27 years. How many years older is Wendy than her niece?

Neutral

Wendy got 4 times as many points on an exam than Dan got. Wendy also got 6 more points than her best friend, Rachel. Rachel wants to get a gold star by earning 15 more points. Rachel's friend got a gold star by getting 2 times the number of points Rachel did. The teacher gave everyone 5 more points. After that Wendy had 3 times as many points as Dan. Rachel didn't earn her gold star. How many more points than Dan did Wendy score?

Mixture Problem

Appropriate

A chemist prepared mixtures of chemicals for an experiment she wants to perform. From her laboratory, she took two 100 ml jars and one 250 ml jar. She then poured from a container a solution of 20% acid into one of the 100 ml jars. Next she took another container and measured a quantity of 30% acid solution into the second 100 ml jar. Finally, she mixed the contents of the two 100 ml containers to obtain exactly 100 ml of solution in the 250 ml jar. When she analyzed this final solution, she discovered the concentration of acid was 23%. How many ml of the 20% solution did the chemist use to make the final solution?

Neutral

Bob, a partygoer, went to a big party last night and had several drinks of punch. For one drink he took three glasses, two 10 oz. ones and one 25 oz. one. He then poured from a bowl a punch containing 20% fruit juice into one of the 10 oz. glasses. Next he took another punch bowl containing a 30% fruit juice punch and poured it into the second 10 oz. glass. Finally, Bob mixed the contents of the two 10 oz. glasses to obtain exactly 10 oz. of solution in the 25 oz. glass. Bob estimated that 23% of this concoction was fruit juice. If that is correct, how many oz. of the 20% fruit juice punch did Bob use to make his drink?

Motion Problem

Appropriate

Two drivers went to business conferences. George has only worked 3 years and goes to the junior executive conference in San Diego. Peggy, on the other hand, has been working 20 years and so goes to the senior conference in LA. George's conference is 10 miles from a beach and Peggy's is 25 miles from one. George drove his first 20 miles at 55 mph, while Peggy, starting 89 miles away from George, started at 65 mph. George's home is 72 miles from his conference, and Peggy is 100 miles from her conference. Due to traffic, George only averaged 27 mph on his trip. Peggy left her house at exactly the same time for her conference, which is 25 miles from George's. Both drivers reach their conference at the same time. How fast did Peggy drive?

Neutral

Two archers went to the archery range to shoot arrows. Phil has only been here 3 times and uses the novice's target range. Rudy, on the other hand, has been here 20 times and uses the intermediate range. The novice range is 10 m from the clubhouse and the intermediate range is 25 m from it. Phil shoots his first 20 arrows and averages 55 m/s each shot, while Rudy, standing 89 m away from Phil, averages 65 m/s. Phil is standing 72 m from his target, and Rudy is standing 100 m from his. Phil aims his next arrow at his target, and fires at a speed of 27 m/s. Rudy fired one of his arrows at exactly the same time at his target, which is 25 m from Rudy's target. Both arrows reach their target at the same time. How fast did Rudy's arrow fly?

(Appendix continues on following page)

Work Problem

Appropriate

Harry the electrician can install a 100 megawatt transformer in 2 hours. A power line powers the transformer at a rate of 500 watts an hour. The electrician follows that power line to the power plant and installs a new 150 megawatt receiver in 3 hours. Harry's apprentice would take 4 hours to install the 100 megawatt transformer. Harry gets a job from a company to install a 100 megawatt transformer and a 10 megawatt receiver. The company also hires another electrician to install a 25 megawatt relay with a maximum power supply of 100 watts/second and to help install the 10 megawatt receiver. The apprentice helps install the 100 megawatt transformer. How long does it take Harry and his apprentice to install the 100 megawatt transformer?

Neutral

A pair of trout can fill a 100 million gallon pond with their offspring in 2 months. A small stream empties into the pond at a rate of 500 gallons an hour. The trout swim up that stream to its source 150 million gallon lake, which they fill with their offspring in 3 months. A pair of carp would take 4 months to fill the 100 million gallon pond with their offspring. The pairs of trout and carp get caught by an ichthyologist and placed in a 100 million gallon tank which is next to a 10 million gallon tank. The ichthyologist also puts a pair of guppies in a 25 million gallon tank, which is supplied with water from the 10 million gallon tank at 100 gallons/hour. How long does it take the trout and the carp to fill the 100 million gallon tank?

Received December 7, 1994
Revision received March 8, 1995
Accepted May 22, 1995 ■