Problem Content Affects the Categorization and Solutions of Problems

Stephen B. Blessing
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
blessing+@cmu.edu

Brian H. Ross
Beckman Institute
University of Illinois
Urbana, IL 61801
bross@s.psych.uiuc.edu

Abstract

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 its solution. Three experiments examined this role of problem content on the problem categorization and solution of algebra word problems with experienced subjects, by manipulating only the content of the problems. When a problem's content was highly correlated with its deep structure (e.g., a content of cars driving for a distance–time–rate problem), people were able to categorize the problem after seeing a smaller portion of it compared to a baseline with contents uncorrelated to the problem deep structure. In addition, for more complex problems in which irrelevant information had been added, problem solving performance was higher and people showed greater sensitivity to the relevance of the information. When a problem's content suggested a different (inappropriate) type of problem, people required a greater part of the problem to categorize it and were slower and less accurate at solving the problem. These results suggest that content may be influential even for experienced problem solvers.

Introduction

Many researchers have investigated how experienced problem solvers use problem information to access and apply relevant knowledge to solve a problem. A common way of characterizing this relevant knowledge is in terms of problem schemata, knowledge structures that allow the identification of problem types and contain associated procedures for solving such problems. In this paper we examine how experienced problem solvers' performance is influenced by the surface content rather than the deep structure that has been the focus of most previous research. The content effects we observed are interesting not only because of their influence on performance, but also because of what they may tell us about how problem schemata are represented, accessed, and applied.

Chi, Feltovich, and Glaser (1981) showed that novices and experts use different problem aspects when sorting problems. The novices put problems together that shared similar surface content features, such as inclined planes and pulleys. The experts, on the other hand, grouped problems according to the principles used to actually solve the problems, such as conservation of momentum and Newton's Second Law. 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).

However, experts can and do make use of the surface content in some situations. Chi et al. (1981) reported the keywords used by experts are a subset of those used by novices. Hardiman, Dufresne, and Mestre (1989) found that experts' judgments of the similarity of physics problems were based not only problems' deep strcuture, but on their content as well. Both of these studies demonstrate that experts do utilize superficial problem aspects in some situations.

Most directly relevant to the current study, Hinsley, Hayes, and Simon (1977) showed that experienced algebra word problem solvers were able to successfully predict the substance of a problem after hearing only the initial noun phrase of a problem ("A river steamer…"). They also found that once a problem has been categorized on the basis of its content, the problem solver may access more knowledge aid solve the problem.

Why do experienced problem solvers use problem content and what might this tell us about the access and use of relevant problem solving knowledge? Although the content of a problem may seem irrelevant to its solutions, often a problem's surface content is strongly predictive of the underlying problem type. In many domains, problems of a given type may have a typical 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 algebra word problems, this correlation between content and type is so great that most experienced solvers describe the problem types in terms of their typical contents, such as work, motion, interest, mixture, age, and river current problems. Indeed, Mayer (1983) found that certain contents appeared with certain problem types more frequently, and that people were sensitive to these frequencies. It would be strange if experienced solvers did not take advantage of the correlations that occur between contents and problem types in order to access the relevant knowledge.

Little research has examined the effect of content on experienced problem solvers. Such effects may be important in understanding how problem schemata are accessed and applied. The problem's content might affect the access or application of problem schemata. The content is available quickly, so it might provide a fast heuristic means of categorizing the problem or accessing the appropriate schema to aid in later instantiation and solution. In addition, the application of the schema might be facilitated by a typical content allowing easier variable instantiation.

The goal of the present research is to use the effects of content to better understand the access and application of problem solving schemata in experienced subjects. In the first two experiments reported here we begin this work by providing a more thorough investigation of content effects, varying the contents for different problem types and examining the effects for both problem solving and on–line problem categorization. The expected result is that problem content will affect the performance of experienced subjects on these tasks, but of particular interest is the pattern of performance across these tasks. In a final experiment, we extend these results to examine whether the content effects might change with the complexity of the problem. In particular, we investigate whether more complex problems might cause experienced solvers to rely more upon the content.

Experiments 1 – 2

Method

Subjects. To ensure that the subjects had a high degree of math competency and previous math experience , we tested 36 graduates of the Illinois Mathematics and Science Academy (IMSA) who were attending the University of Illinois. While perhaps not algebra experts, they can be considered highly experienced in the subject. They were paid for their participation in the experiment, which lasted about 1 hour. Twelve subjects participated in the first (categorization) experiment, and 24 different subjects in the second (problem solving) experiment.

Materials. We constructed 12 triplets of algebra word problems that had the same deep structure (and solution) but different contents (i.e., cover stories and objects). In Appendix A are a set of "motion" (or distance–rate–time) problems. In addition to motion problems, five other problem types were used: age, interest, mixture, river current, and work. The *appropriate* problems have the content typically associated with the problem type (i.e., the content correlated most strongly with the problem type), the *neutral* problems have a content that is not usually associated with any particular problem type, and the *inappropriate* problems have a content typically associated with a different problem type (in the example, an "interest" problem).

In the categorization experiment, each problem was split into a number of phrases (five to nine phrases per problem). Each phrase was printed on a 8.5" x 1" slip of paper. In the example presented in Appendix A, the phrase boundaries are represented by "\", and occur in corresponding positions accross the matched problems. In the problem solving experiment, each problem was presented as a paragraph at the top of a sheet of paper.

Procedure. In both experiments, each subject received 15 problems, 3 practice and 12 test problems. The 3 practice problems were of different types than the 6 types listed previously. The 12 test problems contained 2 each of the 6 types, comprising 4 in each of the 3 conditions. The problems were randomly ordered for each subject.

In the categorization experiment, subjects were tested individually, and their comments tape recorded for later transcription. After receiving the first phrase of a problem and reading it aloud, the subject was asked three probe questions: a) How would you categorize this problem?; b) What information to you expect in later clauses?; and c) What will the final question will be? After responding to the questions, the subject was given the second phrase of the problem and asked the same three questions. This phrase–probe procedure continued until either no phrases remained or until the subject clearly knew what the problem entailed.

In the problem solving experiment, subjects were tested in small groups of 2 to 6. After every 45 s, the experimenter would call out "Line." The subjects then drew a line across the page below where they had been working and continued work below the line. The subjects had 3 min to work on each problem.

Results

The main result of interest in the *categorization* experiment is the proportion of phrases seen before correctly categorizing a problem in each of the conditions. Both authors independently scored the answers, with the few differences resolved by discussion. The score for each problem was the percentage of the problem read before the subject adequately answered the three probe questions. Subjects required a mean of 0.29 of an appropriate content problem in order to correctly categorize it, 0.55 of a neutral content problem, and 0.79 of an inappropriate content problem, $F(2,11) = 55.42$, $MS_e = 0.208$, $p < .01$. All means are different from one another ($p < .01$ by a Newman Keuls' test).

Another important result from this experiment is whether the problems in the different conditions were correctly categorized by the final phrase, when the whole problem had been read. This measure too shows some effect of content. Subjects 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.

The dependent measures from the *problem solving* experiment were the accuracy and latency with which the subjects solved the problems. The accuracy score for each problem was either 0, 0.25, 0.50, 0.75 or 1 depending on the correctness of solution. The problems were independently scored by two people, with discrepancies adjudicated by a third party. Subjects scored a mean of 0.73 on the appropriate content problems, 0.77 on the neutral content problems, and 0.64 on the inappropriate content problems, $F(2,23) = 3.76$, $MS_e = 0.412$, $p < .05$. The inappropriate content problems differed from both the appropriate and neutral ($p < .05$).

Using the lines the subjects drew across the page at 45 s intervals, the midpoint of the interval in which the subject wrote down the equation that once solved would yield their final answer was recorded for every problem. A score of 3 min was given if the subject never arrived at such an equation. It is important to note that the measure is time to equation, not correct equation. Subjects spent a mean of 1.06 min coming up with the equation on the appropriate problems, 1.19 min on the neutral problems, and 1.63 min on the inappropriate problems, $F(2,23) = 15.38$, $MS_e = 3.95$, $p < .01$, again with the inappropriate content problems differing from both the appropriate and neutral content problems ($p < .05$).

Discussion

The categorization experiment demonstrates that subjects can use a problem's content in order to correctly categorize that problem. The extent to which a problem's content usually occurs with a particular problem type affects the proportion of the problem the subject needed to read before making a correct categorization: the appropriate contents were categorized most quickly and the inappropriate contents least quickly. However, most subjects were able to categorize the problems before the last phrase, the question, of the problem was read, even in the inappropriate content problems.

In the problem solving experiment, we expected this effect of content to carry over in the accuracy and speed at which these problems were solved. However, the appropriate and neutral conditions did not differ in either accuracy or speed, despite a large difference in the categorization task. As mentioned, almost all of the neutral problems were correctly categorized by the end of the problem. Thus, to the extent that this categorization measure is indicative of schema access during problem solving, we might not expect any difference in accuracy. The lack of any difference 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 schema access times would be small anyway.

However, content did affect problem solving. The inappropriate condition performance was less accurate and had longer latencies than the performance in the other two conditions. It is interesting to note that in the categorization experiment, not only did the inappropriate condition lead to the slowest categorization, but in about 25% of the cases there was not a correct categorization even when the whole problem had been read. It is conceivable that the worse performance of the inappropriate condition and the similar performance between the appropriate and neutral conditions reflect schema access effects. That is, once a schema has been accessed, the problem's content plays no further role whatsoever. The inability to access a schema in the inappropriate condition, apparently solely because of content, leads to worse problem solving behavior.

Experiment 3

The previous experiments have shown that subjects can use a problem's content in order to categorize a problem, and that, in at least some cases, content can affect problem solving. One possibility for the lack of any difference in problem solving performance between the appropriate and neutral conditions is that the problems used have been quite simple, so that the appropriate schemata could be accessed and applied in either condition by such experienced subjects. To better understand the role of content in problem solving, it is useful to examine how schemata are accessed and applied to more complex problems. In Experiment 3, we investigate this possibility by using problems that have much additional irrelevant information (see Table 2). The hypothesis is that the appropriate contents will allow the experienced solvers to more easily ascertain the relevance of problem information and lead to better performance.

Method

Subjects. The subjects were 16 graduates of IMSA who had not participated in either Experiment 1 or 2. They were paid for their participation in the experiment, which lasted about 45 min.

Materials. The problems used in this experiment were modified versions of the appropriate and neutral content problems from Experiments 1 and 2. Appendix B contains an example of these problems. Inappropriate content problems were not used. The modification involved the addition of irrelevant information. The information added was not needed in order to solve the problem, but was information that could not simply be discarded because it seemed out of place.

Procedure. The procedure was as in the problem solving experiment.

Results

After conducting the experiment, we found that there were serious wording problems with two problems (one appropriate and one neutral), 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 are included.

These complex materials did show an advantage for the appropriate condition. Subjects scored a mean of 0.66 on the appropriate content problems and 0.58 on the neutral content problems, $t(15) = 2.13$, $p = .05$. This advantage of appropriate content was found for 13 of 16 subjects. The time measure showed only a small, non–significant advantage for the appropriate content problems, with a mean of 1.76 min versus 1.90 min for the neutral content problems ($t(15) = -1.60, p > .1$).

We also collected protocols from two algebra experts solving these problems. One had taught algebra for 6 years, the other for 28. We analyzed these protocols for the statements of relevant aspects of the problem (i.e., those needed for solution) versus irrelevant aspects of the problem. This analysis showed that the experts were more sensitive to the relevant information in the appropriate problems than in the neutral problems. In particular, one can examine what proportion of the protocol statements focus on the relevant versus irrelevant aspects of the problem. Subtracting the percent time spent on irrelevant problem aspect from time spent on relevant problem aspects provides a measure of how focused the problem solver was on solving the problem. For appropriate problems, 55% more of their statements were on relevant problem aspects than irrelevant aspects, while this figure was 42% for the neutral problems. This difference suggests that the experts more readily filtered out the irrevelant information in the appropriate content problems thatn in the neutral content problems.

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 2. Although the problem solvers were experienced in solving algebra word problems, the appropriate content problems were solved more accurately. The problem types were the same as in Experiment 2, suggesting that the differences came about not because the subjects did not know the underlying problem types, but because the additional irrelevant information interfered more with the problem solving for the neutral content condition.

The protocols from the algebra experts were very revealing in how they solved these problems. In many cases, soon after beginning to read the problem statement, they would explicitly try to categorize the problem. This category information was then used in order to determine the relevance of problem statement information and to facilitate solving the problem. For example, after quickly categorizing an interest problem, one expert correctly decided two pieces of information were irrelevant ("I am going to ignore the 7% money earned from the saving account for the moment. And ignore what she did with the dividend for the moment."), and this decision helped in solving the problem.

General Discussion

In these studies, problem content affected the categorization of problems and their solutions. For problem categorization, the appropriate content problems were categorized sooner than the neutral, while the inappropriate content problems were not categorized until much more of the problem had been read. In problem solving with the simple problems, performance was worst, in terms of time and accuracy, when the problem's content was inconsistent with the problem's underlying type (the inappropriate problems), even though the problems were the same structurally as the matching appropriate and neutral content problems. However, no difference in time to solution or accuracy was detected between the appropriate and neutral problems. With more complex problems, the appropriate content did lead to more accurate performance than the neutral content problems. In addition, the follow–up protocol study indicates that when the content was appropriate to the problem type, people are more sensitive to the relevance of problem information than when the content is neutral. Thus, when content and underlying problem type are correlated, experienced problem solvers are able to make use of the correlation in aiding their problem solving.

Although these results provide clear evidence for content effects in experienced problem solvers, they do not allow a determination of whether the effects are due to access alone or whether they might also show some effect on applying the relevant schematic knowledge. Some protocols suggest that the contents are influencing performance throughout the problem solving while instantiating variables, but further research will address this issue in closer detail. Such research is important for understanding whether problem schemata contain contents only as routes of access or whether the variables and relations found in problem schemata are specialized as a function of frequent typical contents.

Appendix A

Example Motion Problem (Experiments 1 and 2)

Appropriate Content

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

Neutral Content

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 Content

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?

Appendix B

Example Motion Problem (Experiment 3)

Appropriate Content

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 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 Content

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?

Acknowledgments

This research was supported by Grant 89–0447 from the Air Force Office of Scientific Research.

References

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.

Hardiman, P. T., Dufresne, R., & Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory & Cognition*, *17 (5)*, 627 – 638.

Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In M. A. Just & P.A. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 89 – 106). Hillsdale, NJ: Erlbaum.

Mayer, R. E. (1983). Memory for algebra story problems. *Journal of Educational Psychology*, *74*, 199 – 216.

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