UNIVERSITY OF ILLINOIS

May 1992

THIS IS TO CERTIFY THAT THE THESIS PREPARED UNDER MY SUPERVISION BY

Stephen B. Blessing

ENTITLED How Content Affects the Categorization and Problem Solving

of Algebra Word Problems

IS APPROVED BY ME AS FULFILLING THIS PART OF THE REQUIREMENTS FOR THE

DEGREE OF Bachelor of Science in

Liberal Arts and Sciences

Instructor in Charge

APPROVED:

HEAD OF DEPARTMENT OF Psychology

HOW CONTENT AFFECTS THE CATEGORIZATION AND PROBLEM SOLVING OF ALGEBRA WORD PROBLEMS

BY

STEPHEN B. BLESSING

THESIS

for the

DEGREE OF BACHELOR OF SCIENCE

IN

LIBERAL ARTS AND SCIENCES

College of Liberal Arts and Sciences
University of Illinois
Urbana, Illinois

1992

How Content Affects the Categorization and Problem Solving of Algebra Word Problems

Stephen B. Blessing

University of Illinois

Running head: WORD PROBLEM CATEGORIZATION

Table of Contents

Abstract

Current research in problem solving suggests that people use previously acquired knowledge to categorize a problem and then to solve that problem. Many studies indicate that novices in a field base their categorization mainly on content features whereas experts in that field use the underlying structure of the problem. Few studies, however, have addressed the question of how problem content affects the categorization and subsequent problem solving performance. The experiments presented here examine this issue in the domain of algebra word problems. Problems were constructed in which the content differed among problems that had the same underlying solution structure. Two different experiments were performed. The first experiment demonstrated that people can categorize problems faster when the content of the problem is consistent with the problem type. The second experiment then tested whether problem content would affect problem solving. Subjects were able to solve the problems more quickly and accurately when content reflected problem type. Also, think–aloud protocols were collected from a few subjects, and some qualitative differences were found between problems with different content but same solution structure.

How Content Affects the Categorization and Problem Solving of Algebra Word Problems

People face problems during all of their waking hours—from what to put on in the morning to what to have for supper at night. These problems range from easy to difficult and take from seconds to a lifetime to solve. Due to this variance, the study of human problem solving is challenging and often not easy. These difficulties can be more easily managed if one focuses only on a subset of questions. In this thesis, I will be concentrating on the categorization and solving of algebra word problems.

Through experience in a domain, people acquire knowledge that facilitates problem solving processes. For example, they ill often decide that a new problem is similar to ones that they have solved previously. This process, called categorization, allows the problem solver to use previously acquired information to solve the current problem. The basis for this categorization, commonly referred to as a schema, may contain useful facts, such as formulae, procedures, and references to other schema, in solving problems of that particular category. For example, a person solving a physics problem may categorize a certain problem as an example of "conservation of momentum," and from that categorization utilize the schema containing such things as the final momentum equals the initial momentum and that momentum is the product of mass and velocity.

Schemata, then, resemble procedures the person has developed to identify and solve problems of a particular type. Contents of particular schemata vary from person to person and a person's schemata will change over time as new information is received (Reimann & Chi, 1989).

The study of categorization and schemata use in problem solving has been a relatively recent endeavor. The problem solving methods relying on these two processes are referred to as strong methods, and so depend greatly on the problem solver's knowledge in a particular field and on the representation of that knowledge. Prior to the study of categorization and schemata, more emphasis was placed on weak problem solving methods. Weak methods differ from strong methods in that they do not depend on the problem solver's knowledge base in any particular domain. Rather, investigators believed problem solvers used these weak methods that would apply to any domain. This being the case, a lot of study was done on the actual search process used to find an answer to a problem, with little attention paid to the representation of the problem solver's knowledge. However, as more research was conducted, representation was found to play a very important role in the problem solving process, and so the areas of research gradually shifted from search to representation.

These early investigators, though, did provide the foundation on which the more recent work is based. Their research on the human problem solving system and its constraints is still quite relevant today. Therefore, a review of that work will be presented here, along with a presentation of weak (search) methods. The discussion will then turn to the more recent work being done in human problem solving.

Background. One of the major works in the field of human problem solving is Newell and Simon (1972). Their work synthesized information about the human processing system and provides the foundation for later works detailing the information processing theory of human problem solving. They discuss the physical constraints of the human problem solver and the parts of the information processing system used in solving problems.

The most important constraint on the human information processor is memory. A human can only keep a few items in working memory, where all problem solving processes leave their inputs and outputs. As a rule of thumb, humans can keep track of seven plus or minus two items in working memory (Miller, 1956). This memory is quick, being able to be written to and read from in only a few milliseconds, but it is only a temporary storage space. The main knowledge base for humans, the long-term memory, is thought to be infinite in capacity. Items retrieved from the long-term memory store take only a few hundred milliseconds to be placed in working memory, but to write information to long-term memory can take several seconds.

The information processing system is composed of different parts, and these parts are invariant across different problem types as well as different people (Simon, 1978). The problem itself is called the *task environment*. Put another way, the task environment consists of the specifications and constraints of the problem statement. The problem solver interacts with the task environment (the problem) and creates a *problem space*, an internal representation of the problem. The problem space consists of a set of interconnected states, including the initial (start) state and the goal (end) state. Each state represents the start state, the end state, or a possible intermediary state the problem solver may reach on the way from the start state to the end state of the problem. To solve a problem, one must find, through some search process, the path of states that leads from the givens to the goal, based on the information stored in each state.

The content of a state is very important, since one must decide the next action to be taken by virtue of the information stored in the state. Content here does not refer to specific domain–dependent knowledge, but rather to the features that each state possesses due to how the problem space was constructed based on the task environment. All the states that make up a problem space should be structured in such a way as to be most conducive to solving the problem at hand. Obviously then, the states should contain only information relevant to the problem, with no extraneous information. This information should be arranged in a sensible, easy to access manner. The overall structure of the states' information aids in the problem solving process. That is, a good structure will help in the search to solve a problem whereas a bad structure will perhaps hinder solving a particular problem.

Newell and Simon, among others, outlined different methods to traverse the problem space. These methods, which would be considered weak methods, did not rely on the domain-specific knowledge content of the states. For example, the problem solver could use an algorithmic search process. Examples of such weak methods are breadth-first search and depth-first search, both borrowed from graph theory. The use of both of these methods would eventually cause the problem solver to visit each state, thus ensuring success (given that the problem has a finite problem space). However, since the goal could be the last state visited, these methods tend to be slow and inefficient.

Perhaps a better traversal method would be to use an heuristic search method. An heuristic does not ensure success, for not every state is guaranteed to be visited. The problem solver decides not to visit some states (even though those states may lead to the goal state) based on various criteria, such as distance from current state to the goal state. For example, in one heuristic method, *means-end analysis*, the problem solver compares the current state to the goal state, and then breaks that bigger problem into smaller subgoals by selecting some operator to apply that will reduce the distance between the two states (Simon, 1978). For example, if I had a problem in which I needed to get from my house in Los Angeles to a friend's house in New York, I would find the greatest distance from my current state to the goal, in this case actually getting from Los Angeles to New York, and then select some operator to reduce that distance, probably using an airplane. I now have subgoals I need to accomplish, such as getting from my house to the airport. By working from the general problem to the specific details, creating subgoals along the way, the problem solver can find a path that will lead from the start state to the goal state. Other weak search methods exist, but these will suffice as examples of how the problem space could be traversed.

The human problem solver must have some sort of directions, or program, that directs and organizes the search process. Newell and Simon put forward a modern idea of a production system, originally conceived by Post in 1943 (Anderson, 1983). A production is a statement with two parts, a condition element and an action element. In other words a production resembles the IF-THEN statements used in many computer programming languages. An example production would be: "If the rate and the time is known, then multiply the two to calculate the distance." A system of individual productions can exist as a cohesive unit, allowing independence among the different sets of these productions in the whole problem solving system. Production systems have been developed that accommodate the constraints and actions of a human problem solver.

As can be seen, earlier research in human problem solving has focused on weak methods of solving problems. The emphasis was much more on the search process than it was on the actual representation of the problem solver's knowledge. More recent work, particularly that done on differences between novices and experts, has shown that the domain knowledge a problem solver possesses influences their problem solving performance, and has also changed some of the earlier notions of the search process.

Novices and Experts. Given that experts are subject to the same physical constraints as novices in problem solving and the same parts (problem solver, task environment, and problem space) of the information processing system, what accounts for the difference one sees between a neophyte in a particular field and a person who has been in the same field for twenty years or more? Clearly, a large difference exists in the way an expert and a novice in a particular domain solve problems. Experts in any particular field can solve problems more efficiently and faster than can a novice just learning that field.

Before starting with the differences, one similarity should be made evident. It appears that experts in a particular field, like the novices, do not have a set of over-all general rules to solve all types of problems, those both in and out of the field (Lesgold, 1988), and nor do they simply have a 'better' memory (Chase & Simon, 1973). That is, if a person is an expert in one field, he or she will not be an expert in a non-related field by virtue of being an expert in the first field. Expertise seems to be a very domain dependent, not independent, phenomenon.

The big difference between novices and experts is the large knowledge base that the expert has to draw upon (Reimann & Chi, 1989). The expert has solved problems, perhaps has done research, and has read a lot about the field in which he or she has expertise, whereas the novice simply does not have that warehouse of knowledge to use. The time and effort the expert has put into the field clearly has big benefits when it comes to solving problems in that field. However, the expert also has a more subtle advantage over the novice.

Experts have organized their large knowledge base into larger units of information (de Groot, 1965; Chase & Simon, 1973). For example, the novice may have three separate items to remember whereas the expert has somehow "chunked" those three separate items into one unit. The advantages of this should be readily apparent, since working memory can only hold a few items of information, and so chunking allows for more information to be in working memory at any one time.

Differences also exist in how novices and experts actually go about solving a problem. Novices generally work backwards from the goal, trying to get to the givens whereas an expert will work forward from the givens to the goal (Reimann & Chi, 1989; Lesgold, 1988). When presented with a problem in a field such as physics, novices will often flip through the current chapter, trying to find equations that contain variables needed to be solved, seeing if they have enough information, and then finding more equations until just one unknown quantity remains. Experts, on the other hand, will look at the givens, and work forward from there, trying to arrive at the goal.

Lastly, experts do not need to try as many of the possible paths from the givens as novices do (Chase & Simon, 1973). That is, experts only look at very few different methods of solving the problem, whereas the novice may look at many. One might think that the expert would look at just as many, if not more, solution paths as the novice, since the expert has all of the past instances stored in a knowledge base that needs to be gone through. Apparently the expert has some mechanism that allows him or her to come relatively quickly to a way of solving a problem.

Categorization. Previously I stated how categorization can be used as an aid in the problem solving process. Novices and experts differ in how they categorize problems and in the kind of information they use to base their categorization (Chi, Feltovich, & Glaser, 1981). Subsequently, the initial categorization then affects the person's solution of a problem, depending on what information that categorization allows the problem solver to access for use in problem solving.

Novices tend to make categorizations based on the superficial aspects of a problem (Chi et al., 1981; Ross, 1989). That is, if the problem is overtly about aircraft flying overhead, the novice will categorize that problem as an aircraft problem, and not, for example, a right-angle problem in which one must take account wind drift while flying and figure the length of the hypotenuse. Novices in any discipline will do this. Physics novices will classify physics problems as inclined plane problems or as pulley problems, because they will base their categorization on the physical objects in the problem. Unfortunately, these superficial aspects of a problem do not necessarily suggest an actual solution method to the problem. Better methods of classifying problems exist, and these are the methods that experts most often use.

Instead of the superficial aspects of a problem, experts will categorize problems based on the "deep structure" of the problem (Chi et al., 1981). That is, experts will consider how the problem would actually be solved in making their categorization. The expert, then, would look at a problem with an inclined plane and classify it not as an inclined plane problem, but perhaps rather as a Newton's Second Law problem or as a conservation of momentum problem. The expert's category, then, suggests a solution method to the problem, whereas the novice's category does not.

As can be readily seen, an expert's categorization is generally more helpful than a novice's categorization. A novice may be able to remember some earlier inclined plane problems (after categorizing the present problem as an inclined plane problem) and a method of solving one of those past problems, but not be able to solve the current problem correctly because the current problem does not have the same deep structure as the remembered ones. However, the expert can categorize the problem as a Newton's Second Law problem and then call forth that schema which will contain a viable way of solving the problem.

Categorization and representation of a problem are intertwined, and so experts and novices represent the same problem differently. For example, the building of the representation differs between expert and novice. Studies have shown that experts do more "qualitative analysis," and use their "physical intuition" before actually retrieving the schema (Chi et al., 1981). As an illustration, experts often draw more diagrams and pictures when solving problems than do novices, indicating more thinking about the problem's structure.

A seminal study done by Hinsley, Hayes, and Simon (1977) looked at how people categorized algebra word problems. Five experiments showed that people utilized those abstractions, or schemata, in solving algebra word problems. The first study showed that people can categorize these problems, and that there was agreement in terms of the categories to which each problem belonged. The second experiment demonstrated that people categorize problems soon after beginning to read the problem, perhaps as soon as after the first phrase. For example, after hearing only the starting noun phrase "A riverboat steamer...," some people who are familiar with these problems could categorize the problem based solely on that initial phrase and then explain what the gist of the unread portion of the problem will be.

The last experiments provided evidence that once a problem has been categorized, the problem solver may use that categorization to access more information to solve the problem. The experiments showed that once a categorization had been made, the problem solver formulated the problem based on information stored in memory. This information accessed by the categorization, the schemata, may arise from seeing problems of that particular type many times and eventually constructing a generalization of that problem type.

Chi, Feltovich, and Glaser (1981) performed similar experiments using physics problems. They found that schemata do differ, as stated previously, in content and organization between novices and experts. Reimann and Chi (1989) provide some explanation for what happens after the categorization of a problem and how a schema is used. Once a problem has been categorized, that categorization triggers a particular schema, held in long term memory. Then the appropriate parts and values of the current problem are placed into predefined slots in the schema. These slots are like placeholders, perhaps containing variables which indicate an acceptable value range, or perhaps a default value. Once the slots are sufficiently filled, the problem is now essentially ready to be solved, efficiently and accurately. For example, once a problem solver has successfully categorized a river current problem, an appropriate equation could be called forth from that schema, and the problem solver will then fill in the known values from the problem into the slots, and solve for the unknown.

None of these studies give a satisfying explanation of how problem content affects categorization and problem solving. That is, what role, if any, do surface features play in the categorization process and subsequently how does that affect problem solving? The Chi et al. paper only examined problems where the problem's content could give useful information about the problem type. The Hinsley et al. study does mention that problem content will affect categorization and subsequent solving, but provides no real performance measures. My project attempts to answer this question of how content affects categorization and problem solving by manipulating the surface structure of algebra word problems.

The Experiments. This project was comprised of two experiments examining the role content played in the categorization and problem solving of algebra word problems. Algebra word problems are often thought of as coming in types (e.g. the "river current" type), and each type is usually associated with a particular content (e.g. river boats on a river). The first experiment looked at whether people categorize problems and if the "appropriateness" of content affects that categorization. "Appropriate" here refers to whether the problem content matches the problem type. If the content gives no information about the problem type, then the problem is "neutral." The Hinsley, Hayes, and Simon study showed that people will categorize problems, but the problems given to their subjects in their categorization study were all of an appropriate nature for that problem type. For example, if the problem was a river current problem, the problem was about a river steamer going upstream between two cities. The category of "river current problem" really gives the method of doing the problem. Also, people label a problem as a "river current problem" merely to use a shorthand method to state that in this problem one must use the equation distance equals rate times time, and that one must take into account the river's current in figuring out the actual rate to use.

As illustration, here is an instance of a river current problem:

> A riverboat steamer, which travels 20 km/hr in still water, sails upstream from New Orleans to Memphis. If the river flows at 3 km/hr and New Orleans is 800 km from Memphis, how long will it take the steamer to get to its destination?

After hearing the phrase, "A riverboat steamer...," the problem solver could categorize it as a river current problem and could then expect to receive information regarding the river current and the distance the river steamer traveled. Indeed, that is some of the information given. However, a problem does not have to be about a river steamer to be a river current–like problem. The problem could start, "An escalator...," and contain information about escalator speed and distance and have an identical solution structure to the river current problem. That would be an example of a neutral problem. The first experiment looked at the categorization of both appropriate problems and neutral problems. People should be able to categorize both sorts of problems, as shown by Hinsely et al. (1977) but using only appropriate problems. However, they should be able to categorize problems with appropriate content faster than the neutral content problems, since content does give some information as to type of problem.

The second experiment examined the solving of these algebra word problems. Two data collection methods were used, one in which protocols were collected, in which the method of formulation, either by schema or sentence–to–equation, was of interest, and one in which the subjects did not talk aloud but were given booklets of the problems, in which performance measures (time and accuracy) were of interest. Again, both appropriate and neutral examples of different problem types were used.

In addition, "inappropriate" problems, where the problem content would suggest a different problem type, were added. For example, the problem's content could be about two people working together, which would suggest a 'work problem,' but the solution structure of the problem would be identical to a 'river current problem.' Both experiment's measures were compared across the appropriate, neutral, and inappropriate problems. Both quantitative and qualitative difference should exist among these conditions since subjects categorize problems faster when the content gives truthful information about problem type (Experiment 1). Subjects should be progressively slower and less accurate when solving appropriate, neutral and inappropriate problems. When not able to categorize a problem quickly, subject may resort to different solution methods when solving neutral and inappropriate problems.

Experiment 1

Method

Subjects. Eight paid subjects were used in this experiment, which lasted about one hour. They were all students at the University of Illinois.

Materials. Twenty-four algebra word problems were created. There were six different problem types (age, interest, motion, mixture, river current, and work), all taken from the Hinsley et al. study. An example of appropriate and neutral problems of all six types is given in Appendix A. Twelve appropriate problems were written, two for each problem type.

For each appropriate problem, a neutral problem was written, matching the appropriate in presentation, such as by number of words and clauses. The problems were then split into two sets of problems, the first set contained one appropriate problem of each type and one neutral problem of each type which had been based on the other appropriate problem. The second set, therefore, had those other six appropriate problems and the six neutral problems that had been matched with the appropriate problems placed in the first set.

Each of the problems was divided into between five and nine clauses, with each clause on a separate slip of paper. For example, one of the two appropriate river current problems was divided:

A riverboat...
...can go downstream...
...from town A to town B...
...at 24 km/hr...
...in 1/4 hour less time...
...than it takes to go upstream...
...from town B to town A...
...at 16 km/hr....
...How far apart are the two towns?

The corresponding neutral problem (seen by other subjects) was:

A trolley...
...can go downhill...
...from station A to station B...
...at 24 mph...
...in 1/4 hour less time...
...than it takes to go uphill...
...from station B to station A...

...at 16 mph....
...How far apart are the two stations?

Procedure. The comments made by the experimenter and subject were tape–recorded, to aid in the scoring process. The experimenter gave the subject one phrase at a time and then asked the subject to read the phrase out loud. After the subject had done so for each clause, the experimenter asked the subject if he or she knew what type of problem it was and what additional information the subject expected to receive. The experimenter then gave the subject the next clause, and all earlier clauses were kept in view of the subject. Also, before the final phrase was given, the subject was asked what question he or she expected to be asked.

Each subject had one warm–up problem given in clauses, either an appropriate or a neutral rectangle problem. A 'rectangle' problem is unlike the six other problem types. Half of the subjects were then shown all of the problems in the first set, and the other half received all the problems from the second set. Therefore, each subject was tested on 12 problems, one appropriate and one neutral from each of the six types. Each subject saw all six of the problem types before they saw a second problem of a previously seen type. In this way, each problem from the set of problems had four observations from different subjects.

Measure. The measure used was the number of clauses needed for the subject to categorize the problem. An adequate categorization would be one that included the usual problem category associated with the problem, an almost complete list of the information contained in subsequent clauses, and possibly a prediction of the question to be asked in the problem. Two people scored the transcribed protocol, the experimenter and a research assistant. The subject was recorded as having categorized the problem after they had either explicitly stated the category label or a synonym, explained the structure of the problem, or mentioned the equations used in the problem. The score given for each problem was the number of clauses the subject needed to give such information. If the subject never correctly categorized the problem, their score was the total number of clauses for that problem. The experimenter scored more liberally, basing some of the score on end questions where the subjects explained some of their responses. The research assistant, on the other hand, took a conservative stance, as he did not use the end conversations between the experimenter and subject.

Results

The prediction was that the subjects would be able to categorize the appropriate problems in significantly fewer clauses than the neutral problems, since the content should serve as a cue. As scored by the experimenter, the subject was able to categorize a problem, on the average, after seeing 25.3% of an appropriate problem or 50.3% of a neutral problem, which was a reliable difference by subject (sign test 8 – 0; $t(7) = 11.23$, $p < .001$). A similar result was obtained by the other scorer, the subject being able to categorize an appropriate problem after seeing 40.0%, or a neutral problem after having read 61.5% of the problem (sign test 7 – 1; $\underline{t}(7) = 4.47$, $\underline{p} < .01$).

While it is true that subjects were able to categorize an appropriate problem with less information, a separate question would be if subjects were able to categorize these problems, especially the neutral, at all. For the experimenter's scoring, only 1 appropriate problem out of 48 was not categorized before the final clause, and 5 of 48 neutral problems were not categorized by then. For the other scorer, 12 of 48 of the appropriate problems and 14 of 48 of the neutral problems were not categorized before the question was given. Even though the appropriate problems were categorized faster, subjects were still able to place similar numbers of appropriate and neutral problems into the correct categories before reading the last clause of a problem.

Discussion

From these results, one can see that people do categorize problems earlier when the problem's content reflects the problem's type. If the content of the problem gives no information about the problem type, more of the problem needs to be examined before one can give detailed information about the problem type. The question now arises, does this faster realization of problem type affect how one solves the problem? That is, does expecting certain information and predicting the final question change the method, accuracy, or time in working these problems to solution. Experiment 2 was designed to answer this question.

Because of the result found in this experiment that people are generally able to categorize a similar number of appropriate and neutral problems before the final clause, a set of inappropriate problems were created. With these problems, where the problem's content indicates another problem type, it is speculated that the subject will not be able to classify the problem until sometime after the question, and perhaps will classify the problem incorrectly. This will provide a better comparison of how categorization affects problem solving than if only appropriate and neutral problem were used.

Experiment 2a

Experiment 2 was divided into two parts, 2a and 2b. In Experiment 2a, the subjects were given booklets of problems to solve using pencil and paper. Experiment 2b had subjects think aloud while solving the problems.

Method

Subjects. Twenty-four paid subjects were used in this experiment, which took 45 min to complete. They were all University of Illinois students who are graduates of the Illinois Mathematics and Science Academy (IMSA). IMSA is a state residential high school for gifted students located in Aurora, IL. IMSA graduates were used to assure high overall performance for this experiment.

Materials. The same two sets of algebra word problems were used in this experiment. In addition, one inappropriate problem of each of the six types was added to each set. Each of the two sets now had 18 different problems, one appropriate, one neutral and one inappropriate problem from each of the six problem types. The inappropriate problems were matched as closely as possible to the corresponding appropriate problem in that set. For example, the appropriate river current problem cited earlier ("A riverboat can go downstream from town A to town B at 24 km/hr in 1/4 hour less time than it takes to go upstream from town B to town A at 16 km/hr. How far apart are the two towns?") was rewritten as an inappropriate problem with a work content as: "Jim and Pete work together, but Jim is a faster worker. Jim works at 24 pieces per hour and can finish his standard quota of pieces in 1/4 less hour than Pete can working at 16 pieces per hour. How many pieces are in a standard quota?" Appendix B lists more examples of inappropriate problems. These inappropriate problems were written so that they matched as close as possible to the appropriate problem on which they were based. Table 1 presents summary statistics for the appropriate, neutral, and inappropriate problems on such objective measures as number of lines and words.

Insert Table 1

about here

Procedure. Each subject received a 15 page booklet with a problem on each page. The subjects had 3 min to complete the problem. After every 45 s, the experimenter would call out, "Line," and the subjects would draw a line across their page and continue work below that line. In this way, the three minute interval was divided up into four 45 s intervals. Thus, the time spent on the different stages of each problem could be inferred from the spatial location of the lines. At the beginning were three warm–up problems (of rectangle, right–angle, and probability type; again, these three types differ markedly from the six main types). Half the subjects then had 12 problems from the first set and the other half had 12 problems from the second set. Each subject had two problems from each problem type, and had four appropriate problems, four neutral problems, and four inappropriate problems. Each subject saw one of each problem type before any repetition of problem type. Since each subject did not see every problem from a particular set, the problems types were counterbalanced across subjects. Each problem had eight different observations.

Measures. The time to solve the problem was obtained by using the lines drawn by the subjects every 45 s. In this way, each problem had four time intervals (0 – 45 s, 45 – 90 s, 90 – 135 s, and 135 – 180 s). Once the subject had written down the necessary relations of the problem and had reduced the problem to the one equation that once solved would yield the answer, it was during that interval of the 'necessary equation' that the subject was scored as solving the problem. If the subject thought they had solved the problem, but had actually solved it incorrectly, they were timed as solving the problem during the interval in which they wrote down the equation they used in coming up with the incorrect answer. For example, if the subject wrote down their main equation before the first line (that is, before having worked on the problem for 45 s), they received a 1 for that problem, if the equation appeared between the second and third line (that is, having worked on the problem between 90 and 135 s), they received a 3. If the subject never wrote down an equation that they used to find a final answer, that subject received a 5 (only 17 5's were given, which accounted for 5.9% of the problems). In this way, one could average the time values scored for a particular type of problem and multiply by 0.75 min to obtain a rough estimate of how many minutes on the average a person worked on those problems.

A scale 0 – 1 was used for accuracy. If the subject solved the problem correctly, a 1 was given. If the subject's solution method was entirely wrong, that solution received a 0. Partial credit was given. If the subject made a conceptual error, such as reversing correspondences, a 0.5 was given. A 0.75 was given if the subject made a mathematical error, and a 0.25 if one or more conceptual errors and math errors were made, but the subject still had some idea of how the problem should be solved. Several other methods of scoring were implemented, such as an objective method where the solution received a 1 if totally correct, a 0 if not, and similar results were obtained with each method, and these will be presented as well.

Each problem was scored twice for accuracy, once by the experimenter and once by a research assistant. The few discrepancies, only 12 out of 288, were adjudicated by a third party.

Results

With the results from Experiment 1, a time difference between the three problem types is expected, with the appropriate problems being solved faster than the neutral problems, and the inappropriate problems taking the longest time to solve. Table 2 provides the time data. Since each subject had two problems of each type and had all six problem types before repeating a type, it may also be interesting to note performance on the first half of the problem set, where having seen a problem of the same type before is not an issue. The analysis of variance indicated a significant difference in time, both in the full set measures and in the first half (full set, $\underline{F}(2,23) = 14.71$, $\underline{p} < .001$; first half, $\underline{F}(2,23) = 13.33$, $\underline{p} < .001$). A Newman–Keuls' test on both the full set or first half data, revealed that the differences between the inappropriate and appropriate means and between the inappropriate and neutral means were significant ($\underline{p} < .01$), but that the difference between the neutral and appropriate means was not significant.

Insert Table 2

about here

A difference in the accuracy measures is expected between the appropriate, neutral and inappropriate problems, with subjects doing best on the appropriate problems, and worst on the inappropriate problems. Table 3 shows the results on the full set of problems and Table 4 shows the results on just the first half. These tables present two scoring measures, one using the partial scoring method described earlier and the other giving no partial credit (that is, a 1 was given if the problem was solved correctly, a 0 otherwise). Appendix C gives the data for three other scoring methods. As can be seen, the results among the different scoring methods showed the same pattern. Therefore, the remaining statistics will only involve the standard scoring. The analysis of variance for the accuracy (using the standard scoring) indicated a significant difference in accuracy, again for both the full and first half sets (full set, $F(2,23) = 4.26$, $p < .05$; first half, $F(2,23) = 3.76$, $p < .05$). Using a Newman Keuls' test, the data show a significant difference ($p < .05$) in the neutral and inappropriate means in the full set and a significant difference ($p < .05$) in the differences between the appropriate and inappropriate means, and the neutral and inappropriate means in the first half set. All other pairs in both sets were non–significant by the Newman Keuls' test.

Insert Tables 3 and 4

about here

Discussion

The results of this experiment were largely as expected, with subjects being able to solve the appropriate and neutral problems substantially more quickly and accurately than the inappropriate problems. This result would seem to indicate that the subjects perhaps classified the inappropriate problems incorrectly or not at all, and this interfered with their performance. With the appropriate and neutral problems, on the other hand, where most subjects probably had a good idea of what sort of problem they had by the end of reading the problem, they seemed to be able to use that category information in helping solve the problem as seen by performance gains.

This experiment provided some evidence that people do use category information in helping to solve algebra word problems. Making the assumption that subjects did not correctly categorize the inappropriate problems while reading the problems, the ability to categorize a problem while reading greatly enhances the performance of the problem solver. The last experiment, where the subject thinks aloud while solving the problems, was designed to see if an obvious qualitative difference exists in the way people solve these problems, in hopes of shedding more light on the quantitative results found in this experiment.

Experiment 2b

Method

Subjects. Six paid subjects were used in this experiment, which lasted for about one hour. They were all IMSA graduates now attending the University of Illinois.

Materials. The same materials were used as in Experiment 2a.

Procedure. Each subject was asked to think aloud while solving the problems and being tape recorded. All the subjects had the three warm-up problems from Experiment 2a, and then half the subjects received six problems from the first set of problems and the other half received six problems from the other set. Each problem had one observation. The problems were again counter-balanced in a similar fashion as in Experiment 2a.

Measure. Each problem solution was measured for how long it took the subject to solve and was also scored for accuracy using the partial scoring technique of the previous experiment. The protocol for each problem was examined to determine how the subject arrived at their answer. Subjects used different methods for finding their answers on various problems. Sometimes the subject could go directly from reading the problem to writing down the one necessary equation to solve the problem. In such cases, the subject was scored as using a schema to solve the problem. For some problems, subjects used a sentence-to-equation method, where they formulated each sentence in the problem and then combined them to obtain the one necessary equation. Sometimes, the subject might have used a hybrid of these two mehtods, and sometimes an entirely different method, and these problems were score accordingly.

Results

The averages for the two objective measures in each condition are shown in Table 5. The time measure was taken from the point when the subject began reading the problem to when their final answer was found. The accuracy measure was scored using the partial credit method.

Insert Table 5

about here

Discussion

As can be seen from the objective statistics, subjects were able to find their answer to the appropriate and neutral problems faster than problems with inappropriate content. However, there was no difference in accuracy.

Of more interest here, however, are the methods used by the subjects when solving these problems and thinking aloud. For almost all appropriate problems (11 of 12 in a preliminarv scoring), the subjects seemed to recognize the problem type and apply stored knowledge to solve the current problem. The subjects were able to go directly from reading the problem to setting up the necessary equations, often starting with just the one needed equation. For example, one subject midway through a problem about trains said, "I always hate these rate problems," and then promptly set up the equation $d = r \times t$, the needed equation to solve the problem.

The method used to solve neutral problems seemed to be split evenly (six and six) among using a schema and a brute force, sentence–to–equation method. Promptly after reading an interest problem involving rabbits (where interest problems are usually about money and banks), one subject was able to formulate the problem in one step and then quickly solve the problem from there. On the other hand, one subject after reflecting on an age–type problem about squirrels collecting acorns, wrote down a four equation, three unknown system of equations and proceeded laboriously to solve the problem from there. Usually these problems are solved using only two equations and one unknown.

Some subjects did apparently use a schema for some of the inappropriate problems (5 cases out of 12). It appeared that on these problems, the subject was able to ignore the content and was then able to 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 average. As such, the word 'average' is often in the problem. This was the case in the inappropriate mixture problems (whose content had birthday parties and different ages), and the subject used that. However, on a number of occasions subjects apparently mis–classified a problem early in reading and that adversely affected their problem solving performance. As an example, one subject read the beginning of an interest–type problem with a motion content and said, "...[H]ey, 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. However, for this interest problem, this solver's thought did not help in the solving of the problem, which he ended up not solving correctly.

General Discussion

Experiment 1 demonstrated that people can categorize problems, that they agree to a large extent what category a particular problem falls into, and that the content of a problem affects the categorization of that problem. It is that last point which is of most importance, since prior studies, such as Hinsely et al., have perhaps hinted at such a finding, but have not shown it. Here, content is shown to have a large impact on the speed in which a person can categorize a problem. Almost always, once a problem has been categorized, that category will suggest a procedure to solve that problem. For example, by their nature river current problems require the solver to take into account a constant rate of motion, either added or subtracted, and so an equation almost always used is:

distance traveled upstream = (rate of riverboat - rate of current) × time

Therefore, once a problem has been correctly categorized as a river current problem, problem solvers can then concentrate their effort into finding the values for the variables in the equation. This will of course facilitate the problem solving process, and an initial mis–categorization will hinder the problem solving process.

Experiments 2a and 2b were designed to ascertain if categorization does affect problem solving that this facilitation or hindrance is indeed true, for it may not necessarily be so. That is, the problem solver may categorize a problem based upon the problem's content, but then may not use that category information in actually solving the problem. The problem solver may resort to using a sentence–to–equation method for every algebra word problem, regardless of category. Based on the results of Experiment 2a, it appears that problem solvers do use that category information in solving the problem. The disparity in time to solve and accuracy measures between problems where the content agrees with what type of problem it is (appropriate problems; for example, a river current problem actually about a riverboat) versus problems where the content gives no category information (neutral problems; for example, a river current problem about an escalator rider) versus problems where the content would suggest a different type (inappropriate problems; for example, a river current problem about two workers) point to such a claim. Experiment 2b affirms this result, by showing use of schemata triggered by categorization versus a sentence–to–equation method when unable to categorize while solving these problems and thinking aloud.

These results are similar to those of Hinsley, Hayes, and Simon (1978), in that they show subjects do recognize problem categories, they can in many cases recognize a problem's category early in reading the problem, they have information about the problem categories which is useful for formulating problems for solution, and that they can and often do use this information in solving algebra word problems when their instructions are simply to solve the problems and not in any way call special attention to problem classification. In addition, these results emphasize the importance of content in the categorization and solving of problems. Content influences the speed at which one can categorizae a problem, and also the speed, accuracy, and method one uses while solving the problem.

The question that now must be asked is, what causes a person to categorize a particular problem as a certain type? As mentioned previously, novices often make their categorization decision based on the superficial aspects, like content, of the problem. Experts, on the other hand, will base their categorization on the problem's deep structure, what type of problem it is (Chi et al., 1981). However, the subjects used in Experiment 2a and 2b had some degree of mathematical expertise, yet their performance was affected by the superficial aspects of the problem. Another study (Hardiman, Durfresne, & Mestre, 1989) showed a result to my finding, with experts being adversely affected by a problem's content, but this time in a physics domain. Perhaps even experts base some part of their categorization on these surface features. In practice this may be a useful source of information, since the surface features of a problem are readily apparent, and are often predictive of certain categories, and so can be quickly analyzed and acted upon. Realizing that certain "types" of problems exist and that problems that fall into a particular "type" often share the same content can be a major asset to a problem solver.

A lot more work still needs to be done in this area. Experiment 1, the study with the clauses, was run without any inappropriate problems. It would be interesting to see how a subject would respond to these problems. Perhaps they would initially mis–categorize them as the content would dictate, and then become confused as to the category of the problem by the time the question was asked. More than likely, they would often be taken aback by the question, almost assuredly more so than with either the appropriate or neutral problems.

One might have expected more of a difference between the appropriate and neutral problems. If a schema really does facilitate in solving problems, the quicker one categorizes a problem and thus has access to that schema, performance should be improved. The results here do not support such a claim. However, as pointed out before, it appears most people had correctly categorized the problem before they finished reading it. This would account for the similarity in performance on these appropriate and neutral problems.

It might be possible to show a facilitation in problem solving due to early categorization and schema access. Problems could be written with a particular content, about riverboats for example, that contained enough information to solve two different questions, one being a question normally asked in river current problems, such as rate of the current, and the other a question normally associated with another type of problem, such as having two riverboats work together. In the experiment, one could then ask one subject the first question and another the second question and then look at performance. Or the question could initially not be given to the subject, and the subject is told to solve for whatever they could before being given the question.

To conclude, these experiments attempted to answer two major questions: (a) Does content affect the way a person categorizes a problem, and (b) if categorization is affected, is problem solving also affected, in either quantitative or qualitative ways. In response to the first question, Experiment 1 demonstrated people can categorize problems and that the problem's content affects the speed at which they can do so. Experiments 2a and 2b answered the second question by showing that a person's problem solving ability is also affected by the problem's content, as evidenced by both objective (time and accuracy) and subjective (method of solution) measures.

References

Anderson, J. R. (1983). Production systems in ACT. The Architecture of Cognition. Cambridge, MA: Harvard University Press.

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

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

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

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 processesing comprehension. Hillsdale, NJ: Erlbaum

Lesgold, A. (1988). Problem solving. In R. J. Sternberg & E. E. Smith (Eds.), The psychology of human thought (pp. 188 - 213). Cambridge: Cambridge University Press.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits in our capacity for processing information. Psychological Review, 63, 81-97.

Newell, A., & Simon, H. A. (1972). The theory of human problem solving. In Human problem solving (pp. 787 - 809). Englewood Cliffs, NJ: Prentice - Hall.

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.

Ross, B. H. (1989). Remindings in learning and instruction. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 438 - 469). Cambridge: Cambridge University Press.

Simon, H. A. (1978). Information-processing theory of human problem solving. In W. K. Estes (Ed.), Handbook of learning and cognitive processes (Vol. 5, pp. 271 - 295). Hillsdale, NJ: Erlbaum.

Author Notes

This research was conducted in order to complete the psychology honors program at the University of Illinois. I would like to thank the director of the program, Dr. Michael Coles, and my advisor, Dr. Brian Ross, for his time and assistance in doing this project. The author of this paper is anticipating his start in the fall as a psychology graduate student at Carnegie Mellon. Reprints of this article will not be available.

Table 1

Structural Summary Statistics for Appropriate, Neutral, and Inappropriate Problems

	Lines			Sentences			Words			Syllables		
	A	N	I	A	N	I	A	N	I	A	N	I
Average	3.58	3.67	3.75	4.00	4.08	4.00	44.3	49.3	46.9	65.3	68.2	63.7
St. Dev.	0.90	0.78	0.97	1.04	0.90	0.85	9.9	9.4	14.9	19.3	15.0	20.7
Median	4.0	3.5	3.5	4.0	4.0	4.0	45.5	50.5	40.5	68.0	66.0	59.0

Table 2

Mean Time in Minutes to Solve a Problem in the Appropriate, Neutral, and Inappropriate Conditions

	Appropriate	Neutral	Inappropriate
Full Set	1.46	1.58	2.07
First Half	1.42	1.42	2.23

Table 3

Mean Score per Problem in the Appropriate, Neutral, and Inappropriate Conditions in the Full Set

	Appropriate	Neutral	Inappropriate
Partial	0.73	0.77	0.64
No Partial[a]	0.60	0.60	0.51

[a] The subject received either a 1 for a totally correct solution or a 0.

Table 4

Mean Score per Problem in the Appropriate, Neutral, and Inappropriate Conditions in the First Half

	Appropriate	Neutral	Inappropriate
Partial	0.75	0.78	0.60
No Partial[a]	0.63	0.63	0.48

[a] The subject received either a 1 for a totally cor ːt solution or a 0.

Table 5

Objective Measures from the Protocols

	Appropriate	Neutral	Inappropriate
Time (in min)	1.5	1.7	2.2
Accuracy	0.85	0.75	0.79

Appendix A

Listed below are half of the appropriate problems along with their associated neutral problems used in Experiment 1.

Age

Appropriate. Ann is 2/3 the age of her sister Jill. In 10 years, Ann will be 4/5 Jill's age. How old will the girls be in 10 years?

Neutral. Yesterday Ricky Raccoon had 2/3 as many acorns as Sammy Squirrel. Today they each collected 10 more acorns. Now Ricky has 4/5 as many acorns as Sammy. How many acorns do each of them have?

Interest

Appropriate. Helen has some money and decided to put it in the bank. She put 2/3 of her money into a savings account that pays 9% annual interest. She put the remaining 1/3 into a T–Bill account which pays 15% annual interest. At the end of 1 year she had earned $145 in interest. How much money did Helen initially put in the bank?

Neutral. Gina has some flower bulbs and decided to plant a flower bed. Two–thirds of the flowers she planted were roses that increase in number by 9% each year. The remaining 1/3 of the flowers she planted were tulips that increase in number by 15% each year. At the end of 1 year she had 145 new flowers. How many flowers did Gina initially plant?

Mixture

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?

Motion

Appropriate. A train leaves New York headed for Chicago traveling at a rate of 60 mph. Two hours later, a second train leaves New York headed for Chicago traveling at 90 mph. How long will it take the second train to catch up with the first?

Neutral. A football lineman picks up a fumble and heads for the end zone at 6 yds/sec. Two seconds later, a linebacker takes off after him at 9 yds/sec. How long will it take the linebacker to catch up with the lineman?

River Current

Appropriate. 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?

Neutral. Bill, who frequently hurries on escalators, walks down the down escalator a distance of 30 ft. In an equal amount of time Bill walks up the down escalator a distance of only 20 ft. Bill is capable of walking on a standard staircase at a rate of 5 ft/sec. What is the rate of the escalator?

Work

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?

Appendix B

Listed below are the inappropriate problems used in Experiments 2a and 2b which were based on the appropriate problems listed in Appendix A.

Age (content suggests Mixture)

Inappropriate. A mason mixed 2/3 as much cement in one container as another. He adds 10 liters of cement to each mixer. Now the first has 4/5 the cement as the second. How much does each contain now?

Interest (content suggests Motion)

Inappropriate. A chauffeur has driven 2 limos 1500 miles total. One month he drives Limo A 9% more miles than previously and Limo B 15% more. He drives 183 miles that month. How many miles had he driven each previously?

Mixture (content suggests Age)

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 (content suggests Interest)

Inappropriate. Mary puts some money in a T-bill that earns 6% interest. Two years later Paul puts an equal amount of money in a 9% T-bill. In how many more years will they have earned an equal amount?

River Current (content suggests Work)

Inappropriate. John and two helpers make 30 tiles. Working alone for that time John can make 20 tiles. John and one helper make 5 tiles an hour. What is the rate of a helper?

Work (content suggests River Current)

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 C

Three other scoring methods were also used. In these, no 0.25 or 0.75 scores were given since, in some sense, these two scorings are somewhat subjective. For the Adjusted Up conditions, any score of 0.25 was adjusted to 0.5 and scores of 0.75 were adjusted to 1. In the Adjusted Down conditions, scores of 0.25 were adjusted to 0, and any score of 0.75 was adjusted to 0.5. Finally, in the last condition, Adjusted Middle, all scores of 0.25 or 0.75 were adjusted to 0.5. As can be seen, no matter what method used, the results are similar.

	Appropriate	Neutral	Inappropriate
Full Set			
Adjusted Up	**0.76**	**0.80**	**0.65**
Adjusted Down	**0.70**	**0.74**	**0.63**
Adjusted Middle	**0.73**	**0.76**	**0.64**
First Half			
Adjusted Up	**0.78**	**0.80**	**0.61**
Adjusted Down	**0.72**	**0.76**	**0.59**
Adjusted Middle	**0.76**	**0.78**	**0.60**