The Role of Understanding in Solving Word Problems

DENISE DELLAROSA CUMMINS

Yale University

AND

WALTER KINTSCH, KURT REUSSER, AND RHONDA WEIMER

University of Colorado

Word problems are notoriously difficult to solve. We suggest that much of the difficulty children experience with word problems can be attributed to difficulty in comprehending abstract or ambiguous language. We tested this hypothesis by (1) requiring children to recall problems either before or after solving them, (2) requiring them to generate final questions to incomplete word problems, and (3) modeling performance patterns using a computer simulation. Solution performance was found to be systematically related to recall and question generation performance. Correct solutions were associated with accurate recall of the problem structure and with appropriate question generation. Solution "errors" were found to be correct solutions to miscomprehended problems. Word problems that contained abstract or ambiguous language tended to be miscomprehended more often than those using simpler language, and there was a great deal of systematicity in the way these problems were miscomprehended. Solution error patterns were successfully simulated by manipulating a computer model's language comprehension strategies, as opposed to its knowledge of logical set relations.

Word problems are notoriously difficult to solve. In the study presented here, one type of arithmetic problem was solved by *all* first-grade children when it was presented in numeric format, but by only 29% of the children when it was presented as a word problem. Nationally, children perform 10 to 30% worse on arithmetic word problems than on comparable problems presented in numeric format (Carpenter, Corbitt, Kepner, Linquist, & Reys, 1980). More importantly, as students advance to more sophisticated domains, they continue to find word problems in those domains more difficult to solve than problems presented in symbolic format (e.g., algebraic equations).

This work was supported by National Science Foundation Grant BNS-8309075 to Walter Kintsch and James G. Greeno. We thank Arthur Samuel, Kurt Van Lehn, and an anonymous reviewer for helpful comments on this manuscript. Requests for reprints should be sent to Denise D. Cummins, Psychology Department, University of Arizona, Tucson, AZ 85721.

0010-0285/88 $7.50

This discrepancy in performance on verbal and numeric format problems strongly suggests that factors other than mathematical skill contribute to problem solving success. In this paper, we explore the contribution of a factor we believe to be heavily involved in word problem solving: *text comprehension processes*. We argue that problem texts should be taken seriously as valid discourse entities. Like narratives, word problems require skillful mapping of text input onto the reader's knowledge base if proper comprehension is to be achieved. In the case of narratives, the reader must map linguistic input onto world knowledge concerning (e.g.) actors and their motives. In the case of word problems, the solver must map linguistic input onto knowledge about the problem domain. Nowhere are the ramifications of a breakdown in these mappings more strongly felt than in the domain of children's problem solving, where developing linguistic skills can play havoc with problem solving strategies. Accordingly, we, like many other researchers, have chosen this domain as a starting point for understanding how the nature of problem solving is shaped and colored by a solver's verbal comprehension skills.

Solving Arithmetic Word Problems

Not all word problems are alike. Some problems are much easier to solve than others. For example, even very young children rarely make errors on Combine 1 problems (See Table 1), but frequently make errors on Compare 1 problems. This differential performance changes with age, with performance on these problems becoming nearly equivalent over time. Because problem difficulty patterns change with age, many researchers have adopted a Piagetian view of solution performance characteristics. According to this view, a problem proves troublesome for a child only insofar as the capacities required to process the problem are not yet possessed by the child. While this general view is fairly uncontroversial, researchers disagree as to *which* capacities develop over time to improve solution performance. Explanations generally fall into two camps: those that attribute improved solution performance to the development of logico-mathematical knowledge and those that attribute such improvement to the development of language comprehension skills. We discuss each of these in turn.

The logico-mathematical development view. According to the logico-mathematical explanation of solution difficulty, children fail to solve certain problems because they do not possess the conceptual knowledge required to solve them correctly. Support for the logico-mathematical development explanation was offered by Riley, Greeno, and Heller (1983) and Briars and Larkin (1984).

Riley et al. argue that problem difficulty depends in part on the problem's semantic structure. Nonetheless, they attribute developmental trends in problem solving skill to the acquisition of knowledge concerning logical set relations. To explicate this view, they proposed models of good, medium, and poor problem solving using a schema type formalism. Set knowledge is represented in these models as schemata that specify relations among sets of objects. Their model of good problem solving possesses elaborate schemata that specify high level set relations, such as part–whole, or subset–superset relations. In contrast, their model of poor problem solving ability possesses impoverished schemata that are capable of representing the integrity of individual sets but not their part–whole relations.

Briars and Larkin also proposed a model of problem solving ability that simulates solution performance characteristics. Although somewhat tempered with "set language" and memory resource constraints, the primary mechanisms contributing to solution performance in this model are deficiencies in conceptual knowledge. Unlike Riley et al., however, this conceptual knowledge includes such things as the ability to understand subset equivalences and the ability to understand that things can be undone in time.

The linguistic development view. The linguistic development view holds that certain word problems are difficult to solve *because they employ linguistic forms that do not readily map onto children's existing conceptual knowledge structures*. For example, a child may understand part–whole set relations and yet be uncertain as to how the comparative verbal form (e.g., How many more X's than Y's?) maps onto them. If this were the case, we would say that the child had not yet acquired an *interpretation* for such verbal forms.

Importantly, the linguistic development view implies that word problems that contain certain verbal forms constitute tests of verbal sophistication as well as logico-mathematical knowledge. Accordingly, solution errors on these problems may reflect deficiencies in semantic knowledge, logico-mathematical knowledge, or both. To test the contributions of each, several researchers have manipulated problem wording and observed its effects on solution performance.

For example, consider the following problem:

> There are 5 birds and 3 worms.
> How many more birds are there than worms?

This is a relatively difficult problem for children, with correct performance ranging from 17% for nursery school children to 64% for first graders. The logico-mathematical view holds that this problem is difficult because it requires sophisticated understanding of part–whole relations,

which nursery school children presumably do not yet possess. Hudson (1983), however, reported dramatic improvements in solution performance on this type of problem when the final line was changed to the following:

How many birds won't get a worm?

Correct performance on this version of the problem ranged from 83% for nursery school children to 100% for first graders. Importantly, even nursery school children exhibited sophisticated set knowledge when solving this problem. They did not, for example, simply line up the birds and worms (on an accompanying picture) and count the singletons. Instead, they solved the problem by counting the smaller set (worms) to determine its cardinality, counting out a subset of the larger set (birds) to the same cardinality, and then counting the number of birds remaining and returning that number as the answer. By using this "match–separate" strategy, even nursery school children evidenced a tacit understanding of one-to-one correspondence among sets that possess equivalent cardinality (*subset equivalence*), as well as a sophisticated grasp of *part–whole* set relations. Similar results were found by DeCorte, Verschaffel, and De-Winn (1985), who improved solution performance by manipulating linguistic aspects of problem texts in such a way as to make the semantic relations among sentences clearer. In fact, the influence of problem wording was apparent in the Riley et al. data. For example, mean solution accuracy on Compare 4 (see Table 1) was 25% higher than that on Compare 5 (for second graders), even though these two problems describe identical part–whole set structures, albeit with different words. The same discrepancy was noted for Compare 3 (80%) and Compare 6 (35%), both of which describe the same problem structure with different wordings.

Empirical results such as these are damaging to the logico-mathematical explanation of solution difficulties. If children fail to solve certain problems because they do not possess the conceptual knowledge required to solve them, one would not expect minor wording changes to improve solution performance. Yet this is precisely what is observed. Instead, these results are entirely consistent with the linguistic development view of problem solving development, since they suggest that children find certain problems difficult because they cannot interpret key words and phrases in the problem text.

An unanswered question in this work, however, is just how children *do* interpret the problems they are asked to solve, particularly those that employ troublesome language. This is of some importance because the errors that children make are often counter-intuitive. For example, the most commonly committed error on the birds/worms problem is to return the large number "5" as the answer to the problem. In fact, these "given number errors" constitute a significant proportion of errors committed on word problems (Riley et al., 1983; DeCorte et al., 1985). It is not clear why children believe that the solution to a problem could consist of a number already given in the problem.

In the work to be described here, we attempted to obtain evidence concerning the interpretations that children apply to standard word problems. By doing so, we also tested the two opposing viewpoints concerning solution errors. We accomplished these goals in the following way. In Experiment 1, we required children to recall problem texts either before or after solving them, thereby providing us with information concerning their problem representations. We then compared these recall protocols to solution errors. We predicted that recall errors and solution errors would vary systematically in that solution "errors" would constitute correct solutions to miscomprehended problems. Second, we compared observed error patterns with error patterns obtained when manipulating a computer simulation program's linguistic and logico-mathematical knowledge. We predicted that the best match between the two sets of patterns would be obtained when linguistic knowledge was altered rather than when logico-mathematical knowledge was altered.

In Experiment 2, we tested children's interpretations of word problems by requiring them to generate final questions to incomplete word problems. We reasoned that in order to complete a problem with an appropriate question, it is necessary to understand it properly. We therefore predicted that solution accuracy would vary systematically with question generation performance.

EXPERIMENT 1

In Experiment 1, first grade children were required to recall word problems either before or after solving them. The word problems that were used in this study were the same ones used by Riley et al. (1983). In addition, these children were required to solve the same problems in numeric format. Our hypotheses were as follows: We predicted that solution performance would vary systematically with recall performance, that is, that most of the variance in solution performance would be attributable to comprehension success. Moreover, we predicted that solution errors would in fact be correct solutions to miscomprehended problems, and that miscomprehensions would be primarily in the direction of transformations of difficult problems to easier ones. That is, we expected that when faced with particularly difficult linguistic forms, children would try to simplify them to bring them more in line with linguistic forms with which they were more familiar. Finally, we predicted that the most commonly committed solution errors could be simulated by manipulating verbal comprehension.

Method

Subjects. Thirty-eight first grade children from the Boulder Valley School District served as participants in the study. The children were tested late in the school year (during May).

Apparatus and materials. The 18 story problems used by Riley et al. (1983) served as stimulus materials in the current study. These 18 problems are presented in Table 1. They consist of six instances within each of three major problem types. The problem types are as follows: *Combine* problems, in which a subset or superset must be computed given information about two other sets; *Change* problems, in which a starting set undergoes a transfer-in or transfer-out of items, and the cardinality of the start set, transfer set, or result set must be computed given information about two of the sets; *Compare* problems, in which the cardinality of one set must be computed by comparing the information given about the relative sizes of the other set sizes. The instances within each problem type differ in terms of which set cardinality must be computed and the wording of the problems. The story problems used in the present study all contained "Mary and John" as actors and "marbles" as objects. This was done to reduce the memory load required to comprehend the problem. The child needed only to attend to the relationships among the sets and to remember the numbers stated in the problems.

TABLE 1

Problems Used in Experiment 1 (Adapted from Riley, Greeno, & Heller, 1983)

Combine problems

1. Mary has 3 marbles. John has 5 marbles. How many marbles do they have altogether?
2. Mary and John have some marbles altogether. Mary has 2 marbles. John has 4 marbles. How many marbles do they have altogether?
3. Mary has 4 marbles. John has some marbles. They have 7 marbles altogether. How many marbles does John have?
4. Mary has some marbles. John has 6 marbles. They have 9 marbles altogether. How many marbles does Mary have?
5. Mary and John have 8 marbles altogether. Mary has 7 marbles. How many marbles does John have?
6. Mary and John have 4 marbles altogether. Mary has some marbles. John has 3 marbles. How many does Mary have?

Change problems

1. Mary had 3 marbles. Then John gave her 5 marbles. How many marbles does Mary have now?
2. Mary had 6 marbles. Then she gave 4 marbles to John. How many marbles does Mary have now?
3. Mary had 2 marbles. Then John gave her some marbles. Now Mary has 9 marbles. How many marbles did John give to her?
4. Mary had 8 marbles. Then she gave some marbles to John. Now Mary has 3 marbles. How many marbles did she give to John?
5. Mary had some marbles. Then John gave her 3 marbles. Now Mary has 5 marbles. How many marbles did Mary have in the beginning?
6. Mary had some marbles. Then she gave 2 marbles to John. Now Mary has 6 marbles. How many marbles did she have in the beginning?

Compare problems

1. Mary has 5 marbles. John has 8 marbles. How many marbles does John have more than Mary?
2. Mary has 6 marbles. John has 2 marbles. How many marbles does John have less than Mary?
3. Mary has 3 marbles. John has 4 marbles more than Mary. How many marbles does John have?
4. Mary has 5 marbles. John has 3 marbles less than Mary. How many marbles does John have?
5. Mary has 9 marbles. She has 4 marbles more than John. How many marbles does John have?
6. Mary has 4 marbles. She has 3 marbles less than John. How many marbles does John have?

Each child solved 18 problems. Half of the 18 problems were first solved and then recalled; the remaining half were first recalled and then solved. Two versions of problem presentation were used to ensure that all 18 problems were tested in each solve–recall condition. In the first version, one half of the problems in each problem type was assigned to the Solve–Recall condition, and the remaining halves were assigned to the Recall–Solve condition. In the second version, these assignments were reversed so that version 1 Solve–Recall problems became Recall–Solve problems, and version 2 Recall–Solve problems became Solve–Recall problems. The presentation version served as a between-subjects factor.

The number triples used in the story problems included only the numbers 1 through 9 and were chosen such that correct answers were (a) less than 10 and (b) not the same as a number used in the story. Nine triples that satisfied these constraints were chosen for use in the problems: 3-2-5, 4-2-6, 5-2-7, 6-2-8, 7-2-9, 4-3-7, 5-3-8, 6-3-9, and 5-4-9. Half of these triples were tested as addition problems and half as subtraction. Across subjects, these triples were assigned to problems such that each triple was tested in each of the eighteen problems.

In addition to the story problems, these number triples were tested as numeric format problems. Each child received the same number assignment condition for both the story problem and numeric format tests. For example, a given child received 3 + 2 = ? as both a story problem (Combine 1) and as a numeric format problem. The child's performance on that equation could therefore be observed under both the story and numeric formats. The numeric formats mirrored the story problem structures to which they corresponded. Note that in certain cases (e.g., Change 5) this meant that the equation to be solved contained an unknown on the left side of the equation (e.g., ? + 2 = 5). All numeric format problems were presented in vertical sentence form; equations such as ? + 2 = 5 were written as an open box with "+2" underneath it, a line underneath "+2," and "5" underneath the line.

Procedure. Children were tested individually in a quiet room in their schools during school hours. In keeping with the methodology of Riley et al. (and others), all problems were presented orally, and the child was required to solve them without benefit of paper and pencil. The sessions were recorded on a small, unobtrusive tape recorder. The child was informed of the presence of the tape recorder, but was assured that only the experimenter would hear the tape (i.e., parents and teachers would not). No child seemed uncomfortable having the session taped.

Problem presentation was randomized for each child. The session began with instructions, followed by practice problems. The practice problems consisted of two solve–recall and two recall–solve problems. Children were assisted in solving and recalling these if required. Once the experimenter was satisfied that the child understood the procedure, the experimental session was begun. Children were not assisted in solving or recalling experimental problems. They also were not told whether a problem was to be solved first or recalled first until *after* the problem had been read. This was done to ensure that the strategies used to solve and recall the problems would be the same in both conditions. Following the oral story problem session, the child was given a sheet with the numeric problems on it and was required to solve these.

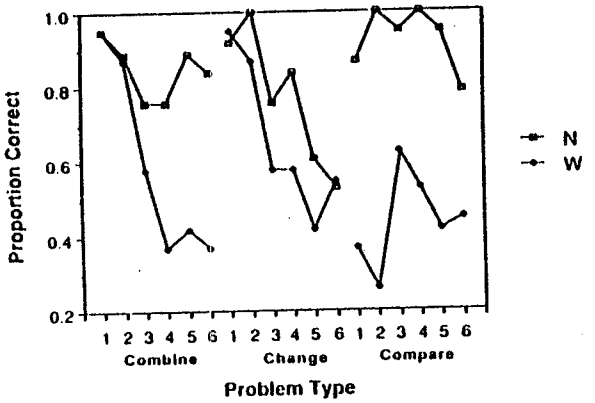

FIG. 1. Proportion of correct solutions for the problems shown in Table 1 when presented as word problems (W) and when presented in numerical form (N).

RESULTS AND DISCUSSION

Protocols were scored for the following: (a) solution accuracy on verbal format problems, (b) solution accuracy on numeric format problems, and (c) structural recall of each problem. The data are pooled over the recall-before and recall-after conditions since initial analyses showed this factor to be nonsignificant. Unless otherwise stated, rejection probability was .05. Significant interactions from ANOVAs were followed by simple effects tests (Keppel, 1973). Significant main effects involving more than one mean were tested using Tukey's test of pairwise comparisons.

Recall and Solution Accuracy

Figure 1 depicts the proportion of subjects who correctly solved each of the 18 word- and numeric-format problems. These results are quite similar to the results of Riley (1981) and Riley et al. (1983), with the possible exception that our students performed slightly higher on the more difficult Change and Compare problems. As expected, performance on numeric problems was consistently higher than that on verbal problems. Some problems were solved correctly more than three times as often in numeric format than in verbal form. Some numeric formats, however, also proved troublesome for the children. These were the number sentences that contained variables, as in "? + 5 = 8" (i.e., Change 3, 4, 5, and 6). First grade children in the Boulder Valley School District are not routinely exposed to number sentences of these forms. Given this fact, it is not surprising that children performed less well on these than on the more typical "3 + 2 = ?." What is surprising is that these number sentences were solved correctly nearly two-thirds of the time, despite their relative novelty.

As stated earlier, subjects' verbal recall protocols were scored for accuracy of *structural recall*. A correct structural recall was any recall that preserved the logical relations among sets in the original problem. For example, consider Compare problems 4 and 5. These two problems describe the same problem structure using different wording. In both cases, the small set must be derived given information about the large and difference sets. "Recalling" Compare 5 as Compare 4, therefore, constitutes accurate structural recall because the original problem's logical structure is preserved. Structure-preserving recall transformations such as these were observed on 12% of the trials; along with veridical reproductions (45%), they constitute our measure of correct structural recall. Together, they constituted 57% of all recall instances.

Figure 2 illustrates proportion correct structural recall. Like the solution results, the recall data are also in agreement with those of Riley (1981), although the greater sensitivity of our recall measure provides a bit more information than the repetition measure used in the Riley study. As predicted, the overall pattern of recall accuracy closely resembled that of word problem solution accuracy, suggesting a strong relationship between the two.

We predicted that word problem performance would vary systematically with recall performance but not with numeric format performance. To test this hypothesis, each subject's protocol was scored for proportion of correct word problem solutions, numeric format solutions, and structural recall across the 18 problems. A regression model was then constructed to predict each subject's overall word problem solution performance as a function of his or her overall performance on the structural recall and numeric format tasks. A forward selection procedure was used

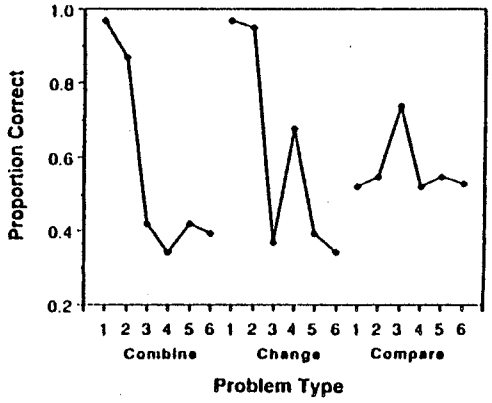

FIG. 2. Proportion correct structural recall for the problems shown in Table 1.

to select candidates for entry in the model.[1] Only one variable met the 5% significance level for entry into the model, that of *structural recall*. This simple model accounted for 72% of the variance in solution accuracy ($F(1,36) = 93.62$, $MS_e = .01$, $p < .0001$), supporting our prediction that performance on word problems depends primarily on successful comprehension.

Finally, to ensure that our observed relationship was not simply reflecting subject variation (i.e., talented subjects performing well on all tasks, less-talented subjects performing poorly on all tasks), we calculated a 2 × 2 contingency table for each subject indicating the number of times problem recall and solution were equivalent in accuracy (i.e., both right or both wrong) or were different in accuracy (i.e., one right, the other wrong). The expected frequency was then computed for the right–right case, and the deviation between observed and expected frequency in that case calculated. These deviations were found to be significantly greater than zero, $t(37) = 6.75$, $p < .001$, indicating that a dependency between solutions and structural recall existed for individual subjects, regardless of talent.

Miscomprehensions and Error Types

While the quantitative relationship between recall and solution performance strongly suggests that solution difficulties are driven by break downs in story comprehension, we can offer more direct evidence by way of qualitative error analyses. We assume that when a child recalls a problem, he or she describes the problem representation he or she constructed during a solution attempt. The nature of a miscomprehension therefore should be related to the type of solution error made. In the following discussion, we will describe the relationships we noted between comprehension and solution errors.

Types of miscomprehensions. Aside from verbatim recall, subjects' recall protocols could be classified into six categories. The first, termed *structure-preserving transformations* (SP), was mentioned in the structural recall analysis and comprised 12% of all recall trials. These were occasions on which the wording of the problem was changed during recall, but the all-important mathematical relations among sets was maintained (e.g., a subtraction Compare 5 problem became a subtraction Compare 4 problem). One interpretation of these transformations is that children are reconstructing the text base from their internal representation of the problem structure, that is, given their comprehension of the logical set relations. If difficult problems are difficult because their text bases are unnecessarily complex, then we would expect children to show evidence of simplifying the text base during recall, that is, transforming a complex subtraction story such as Change 5 into a simpler subtraction story such as Change 2. To test this prediction, we divided each of the three problem types into easy and hard problems based on solution accuracy levels reported by Riley et al. Easy problems included Combine 1, 2 and 3, Change 1, 2 and 3, and Compare 1, 3, and 4. The remaining problems were classified as hard problems. As predicted, more transformations of hard problems into easy problems were observed than vice versa. Of the 25 subjects who produced these recall transformations, 4 exhibited tied scores and 18 exhibited the predicted bias, $p < .01$ via the sign test. Moreover, when transforming a problem during recall, children seemed to be sensitive to problem type. On 77% of these misrecall occasions, the problem was transformed into simpler problems *of the same type*, e.g., difficult Compare problems were transformed into easier Compare problems, difficult Change into easier Change. The only exception was a slight tendency to transform difficult problems of all types into Combine 1, the problem that occurs most often in arithmetic textbooks (Stigler, Fuson, Ham, & Kim, 1986). Combine 1 seems to be a "default" problem, and when all else fails children resort to what is most familiar to them.

The second type of misrecall category involved *structure-violating transformations* (SV). This category comprised 12% of all recall trials. It included occasions in which problems were transformed into other legitimate problems, but the transformation violated the important mathematical relations of the original problem (e.g., a subtraction Compare 5 became an addition Compare 3). In these cases, both wording and structure changed. As in SP transformations, these misrecalls showed a bias toward oversimplification: Of the 21 subjects who produced these transformations, 1 exhibited a tied score and 20 exhibited the predicted bias, $p < .001$ via the sign test.

The third type of misrecall (NP) comprised 8% of all trials and involved recalling the problem as the following *n*onsense *p*roblem:

(a) Mary has 5 marbles.
John has 4 marbles.
How many does (Mary, John) have?
OR
Mary had some marbles.
John gave her 3 more.

[1] Separate regressions were also performed on the three problem types. The results were not appreciably different than the overall regression, with the exception that the regression coefficient for numeric accuracy was marginally significant for Change problems ($b_n = .22$, $p < .11$). This was not surprising since, as noted earlier, four out of these six problems contain unknowns in their number sentences, and first grade children are not familiar with such forms.

Now Mary has 7 marbles.
How many does Mary have now?

These are nonsense problems since they require no computation and instead simply ask for one of the numbers given in the problem.

The fourth category (2S) included recall instances such as the following:

(b) Mary and John have 5 marbles altogether.
Mary has 3 marbles.
How many do they have altogether?

This misrecall contains *two* references to the superset in the problem, once in the body of the problem, and once as a request for the superset cardinality. This is similar to problem category NP in that the question requests a number already given in the problem. Misrecalls such as these were observed on 4% of all trials.

The fifth category (0S), contrasts sharply with category 2S. Instances in this category were observed on 6% of all trials. They consisted (e.g.) of the following:

(c) Mary had some marbles.
John gave her 3 more.
How many did Mary have in the beginning?

Here, the superset specification line (i.e., "Now Mary has 7 marbles.") is simply left out of the problem altogether. In contrast to the other transformations, this category does not seem to be a "transformation" at all, but rather a legitimate "misrecall," or memory error. A line from the problem was simply left out or forgotten. Compare this to category 2S, which seems to suggest a true misconceptualization of the problem structure.

The sixth and final category simply included all misrecall instances that did not fit into the above categories either because the child could remember nothing of the problem, or recall was so confused it could not be classified. This category comprised 13% of all trials.

In summary, subjects' miscomprehensions appeared to be systematic in that they could be classified into five meaningful categories. It is also interesting to note the distributions of problem types across these categories. When Compare problems were miscomprehended in a classifiable way, they tended to fall into two categories, SV (38%) and NP (31%). Change problems also tended to fall into the SV category (40%) and the NP category (20%). Combine problems, on the other hand, tended to be miscomprehended as double-superset problems (33%). Clearly, some aspect of these problems tends to invite certain interpretations from children. We return to this question below; here, we turn to the more important question of how these recall miscomprehensions affected solution performance.

Miscomprehensions and solution performance. Aside from correct solutions (55%), subjects in this study produced wrong operation errors (8%), given number errors (18%), arithmetic errors (11%), and unclassifiable errors (8%). *Wrong operation errors* are errors in which the child used an incorrect arithmetic operation in an attempt to solve the problem, i.e., added when subtraction was required or subtracted when addition was required. *Given number errors* are errors in which the child returned one of the numbers in the problem as the answer to the problem. These types of errors are particularly important to our analysis because they betray faulty comprehension of the problem structure. Moreover, they are frequently reported in the literature (Riley et al., 1983; Decorte et al., 1985) and comprised the majority of errors committed in this study (58%).

We predicted that there would be a systematic relationship between story miscomprehension and solution errors. In particular, we predicted that "errors" would often be "correct answers" to miscomprehended stories. Such relationships were observed; they are presented in Table 2. We discuss this table in some detail below.

Beginning at the top of the table, notice that *structure-preserving transformations* (SP) were most often associated with corrects answers (64%) and arithmetic errors (25%). *Structure-violating transformations* (SV), on the other hand, were most often associated with wrong operation errors (42%). In fact, only 14% of these SV transformations were solved correctly. Looked at from the other side, 64% of ALL wrong operation errors tended to be associated with this type of transformation. Intu-

TABLE 2
Miscomprehensions and Conceptual Solution Errors

Recall type	Response type						Total
	CO	WO	SPN	SBN	AR	OTH	
Structure Preserving (SP)	51	1	5	2	20	1	80
Structure Violating (SV)	11	34	16	4	12	3	80
Nonsense Problem (NP)	10	2	14	24	4	1	55
Double Superset (2S)	4	5	13	0	5	3	30
Partial Recall (0S)	25	1	2	4	5	4	41
Correct (CO)	251	3	11	13	20	13	311
OTH (Other)	25	7	12	5	6	32	87
Total	377	53	73	52	72	57	684

Note. Frequencies are based on 18 observations from each of 38 subjects. WO, Wrong operation errors; SPN, superset given number errors; SBN, subset given number errors; AR, arithmetic error; OTH, unclassifiable error; CO, correct solution. See the text for an explanation of recall types.

itively, these results make sense. Transforming a problem in a structure-violating way meant turning an addition problem into a subtraction problem, and vice versa. It follows, then, that a wrong operation error would occur on these trials.

Nonsense-problem transformations were associated overwhelmingly with given number errors (69%), and in virtually all cases, the number returned was the cardinality of the set specified in the final line of the transformed problem. Correct solutions were observed on only 18% of these trials. Again, these results make sense intuitively. If a problem were transformed into the nonsense problems described above, the transformation would produce a "problem" that requests a given number. Accordingly, a given number is returned as the answer.

Double-superset transformations (2S) were associated with given number errors (43%) and wrong operation errors (16%). All but one of the given number errors involved returning the superset quantity as the answer. Virtually all of the wrong operation errors consisted of incorrectly adding rather than subtracting the numbers in the problem. Correct solutions occurred on 17% of these trials, and the majority of these were on trials in which the original problem was an addition problem as well. The addition of a superset line to the transformed problem therefore seems to be interpreted by children in one of two ways, namely: (1) Add the numbers or (2) return the large number (i.e., return how many there are altogether).

Interestingly, *partial-recall* (0S) transformations were associated most often with correct solutions. On 61% of the trials in which this misrecall occurred, the problem was solved correctly anyway. This type of transformation therefore seems to be a genuine "misrecall," or memory error, as suggested earlier.

In summary, structural recall, both correct and erroneous, provided clear evidence that children's problem solving strategies are determined by their comprehension of the problem stories. Moreover, frequently observed conceptual errors were related to story miscomprehension in systematic ways. These conceptual errors were found to be correct answers to miscomprehended stories. Subtraction tended to be used to solve addition problems that were miscomprehended as subtraction problems, and vice versa. Given number errors tended to occur on trials where problems were miscomprehended as nonsense problems or double-superset problems, that is, as problems that simply request one of the numbers in the problems. Children, therefore, correctly perform operations that they believe are requested by the problem statements.

The more important question, however, is *why* children tend to miscomprehend problems in the ways that they do. As is apparent in our data, these miscomprehensions were not idiosyncratic; recall protocols could be classified in meaningful ways. This suggests that certain aspects of these problems tend to invite interpretations that are similar across subjects. The alternative views we contrast here to explain these misinterpretations are (1) that comprehension failures reflect deficiencies in conceptual knowledge, and (2) that comprehension failures reflect inadequate mappings from English phrases onto existing conceptual knowledge. To test these two views, we employed a computer simulation model.

Model Predictions of Solution Performance

In order to evaluate the relative contributions of logico-mathematical knowledge and linguistic knowledge to solution accuracy, we required a computer simulation model to solve these word problems under four knowledge conditions: full knowledge, impaired logico-mathematical knowledge, and two versions of impaired linguistic knowledge. The simulation is based on a model of children's problem solving proposed by Kintsch and Greeno (1985). A full description of the computer model is given by Dellarosa (1986) and Fletcher (1985). We offer here a rather detailed summary of the model due to its importance in interpreting our results.

The most fundamental aspect of the model is that it solves problems through an interaction of text-comprehension processes and arithmetic problem solving strategies. Its ability to solve them correctly, therefore, depends on the integrity of both types of knowledge.

In solving problems, the computer model first comprehends the story by building proposition frames which represent the story's text base (van Dijk & Kintsch, 1983). Numeric information, however, is given special treatment. Such information is used to build representations of sets, called set frames. In a sense, the simulation "understands" numbers as sets of objects. The local relations among these sets are captured in superschemata, which are larger set frames that have whole sets as their components. A superschema in this model is essentially a blended representation of story and problem structure.

There are three basic types of superschemata in the model, SUPERSET, TRANSFER, and COMPARE. They represent text-driven mappings between part–whole set relations and the story situations described in COMBINE, CHANGE, and COMPARE problems, respectively. For example, SUPERSET is a superschema that contains three sets, two of which are *parts* of the third (e.g., 3 dolls and 2 teddy bears = 5 toys). This is the most transparent mapping. TRANSFER-IN/OUT are superschemata that identify an original set, a set of objects transferred into or out of that set, and a set representing the results of the transfer (e.g., Mary's 5 toys, John's gift to Mary of 2 toys, Mary's resulting 7 toys). This is

essentially a part–part–whole structure that has been contextualized to map smoothly onto CHANGE type problems. COMPARE is a superschema that identifies two sets and the set representing the difference between their cardinalities (e.g., John's 5 toys, Mary's 2 toys, and the 3 toys Mary has more than John). Again, this is essentially a whole–part–part structure that has been contextualized to map smoothly onto COMPARE type problems.

Sets are assigned to the slots in these contextualized schemata based on the relations among propositions in the text base and the sets already existing in short-term memory. For example, if the current line of the text base contains a transfer proposition (i.e., GAVE MARY JOHN (5 MARBLES)), and a set belonging to MARY already exists, then the set already in existence is assigned to the role of the STARTSET and the set referred to in the proposition (i.e., 5 MARBLES) is created and assigned the role of TRANSFERSET. The existence of these two sets and the transfer proposition then triggers the creation of a global TRANSFER-IN superschema.

It is important to note that these schemata are constrained to direct mappings of propositional structures onto logical structures. Essentially, that means that they are constrained to directly mirroring the actions or descriptions in the text base. Certain problems, however, do not lend themselves readily to this proposition-schema mapping; instead, they require that inferences be made concerning abstract part–whole relations. Take, for example, Changes 5 and 6. These problems start out with an unknown quantity. Transfers cannot be made into or out of an unknown quantity. As a result, the simulation (and presumably children) must infer super-set and result-set role assignments given the *complete* story situation. The simulation does this by way of conversion rules. These rules are simply mappings from complete TRANSFER-IN/OUT superschemata (in which the start set is unknown) onto abstract part–whole schemata. Similar conversions are done for Compares 3 through 6 since these problems do not specify direct comparisons between two set quantities but instead require inferences about which set among the three is the superset.

Consequently, the simulation also contains other decontextualized knowledge concerning part–whole relations, in addition to these contextualized schemata. The most important is a decontextualized part–whole superschema. Assignment of sets to this schema is done either through the contextualized conversion rules mentioned above or through decontextualized reasoning strategies. For example, the model assumes that the least specified of three sets of like objects can be assigned the role of WHOLE in a part–whole schema (e.g., windows = big windows and small windows). It can also make assignments based on class membership (e.g., 5 toys = 3 dolls and 2 teddy bears) and on conjunction information (e.g., John & Mary's 7 marbles = John's 5 marbles and Mary's 2 marbles.) In each case a part–whole (or SUPERSET) superschema is created to capture the logical relations among the sets specified in the problems.

Finally, the presence of superschemata triggers arithmetic counting procedures which produce answers to the problems. Failure to produce an adequate superschema (i.e., misunderstanding the problem) causes the program to use default strategies to produce a "best-guess" answer. An example of such a default strategy is to search memory to determine whether the answer is already known, that is, if a set that matches the requested specifications has already been created. Another default strategy is to mine the text base for key words (e.g., altogether, "in the beginning") that might cue a solution procedure. The effect of these default strategies will become apparent shortly.

When given the 18 problems to solve, the Dellarosa (1986) simulation model solved all 18 without error, indicating that it and the Kintsch and Greeno (1985) model upon which it is based are sufficient models of children's problem solving. More germane to our discussion here, however, is its usefulness in explaining children's errors. Specifically, we required the model to attempt these same problems under conditions of impaired knowledge and compared its performance to that of children. Presented in Table 3 are the answers produced by the simulation under each of three knowledge impairment conditions, along with children's errors observed in this study.

Deficient conceptual knowledge. To test the contribution of conceptual knowledge, we removed the simulation's decontextualized knowledge concerning part–whole relations and required it to solve the 18 problems.

The first thing to notice is that without conceptual knowledge, the simulation's performance matches that of children on four problems. Its solution protocols can be described as attempts to model the actions in the story, relying on linguistic knowledge and default strategies to obtain answers. Relying on its "altogether means add" key word strategy, it solves the addition Combine problems correctly (Combines 1 and 2), but produces wrong operation errors on the subtraction Combine problems (Combines 3 through 6). Children, however, produced given number errors most frequently on these problems, with the exception of Combine 5. On this problem, both simulation and children produced wrong operations errors. Also, like children, the simulation had little difficulty solving Change 1 and 2 problems because these two describe only simple transfer operations. It cannot solve Changes 3 through 6, however, because the story actions describe transfers involving unknown quantities, and it has no way of mapping these onto part–whole structures. It cannot resort to its default strategy of simply returning the quantity of the set specified in the last line of the problem because that quantity is "SOME," and that is

TABLE 3
Characteristic Errors: Observed and Simulated—Experiment 1

Problem	Children's errors (most frequent)	Simulation error		
		(¯C)SL	C(¯S)L	CS¯L
Combine 1	—	—*	—*	—*
Combine 2	—	—*	—*	—*
Combine 3	SPN	WO	—WO	SPN*
Combine 4	SPN	WO	—WO	SPN*
Combine 5	WO	WO*	—WO*	SPN
Combine 6	SPN	WO	—WO	SPN*
Change 1	—	—*	U	—*
Change 2	—	—*	U	—*
Change 3	SPN	U	—	—
Change 4	SPN	U	—	—
Change 5	SPN–SBN	U	—	SPN–SBN*
Change 6	SPN–SBN	U	—	SPN–SBN*
Compare 1	SPN–WO	—	WO*	SPN*
Compare 2	SBN–WO	—	WO*	SBN*
Compare 3	SBN–WO	U	WO*	SBN*
Compare 4	SBN	U	—	SBN*
Compare 5	SBN	U	—	SBN*
Compare 6	WO	U	WO*	SBN
Total matches (*)		5	7	14

Note. Children's errors constitute the most frequently observed error based on 38 subjects' observations per problem. Simulation conditions: ¯CSL, no conceptual knowledge; C¯SL, no problem situation knowledge; CS¯L, degraded linguistic knowledge concerning key words and phrases. Error types: WO, wrong operation errors; SPN, superset given number errors; SBN, subset given number errors; —, no error; U, unable to derive any solution.

* Problems for which the simulation matched children's error patterns.

not an acceptable answer. As a result, it produces no answer at all, unlike children who tended to produce given number errors on these problems. Finally, it has no difficulty with Compares 1 and 2 because it can perform the simple comparison of quantities required by those problems. Note, however, that these are two of the most difficult problems for children to solve. It cannot solve Compares 4 through 6 because they require mappings onto abstract part–whole structures in order to determine whether addition or subtraction is in order.

Deficient story-situation knowledge. To test the contribution of story-comprehension components, we removed the schemata that correspond to Combine, Change, and Compare problems and restored its decontextualized conceptual knowledge. In other words, we removed the contextualized mappings from *story situations* to part–whole structures. This means that the simulation could not understand whole story situations,

but instead could only search for key words and the like in order to trigger its conceptual knowledge concerning part–whole relations. Under these conditions, the simulation matched children's response patterns on 7 out of 18 problems. Its solution protocols revealed the following:

Using its "altogether" default key word strategy, the simulation successfully solved Combine 1 and 2 problems. Its performance on the other Combine problems, however, depended on the ordering of rules/strategies. If the rules were ordered such that the simulation accessed its default rules rather than its thinking rules, then it committed wrong operation errors on Combines 3 through 6. If the rules were ordered such that the simulation "thought" before "defaulting," then it used its conjunction-superset strategy to assign the role of SUPERSET to the set owned by both Mary and John. As a result it solved these problems correctly. Giving priority to defaulting produced a pattern that matched children's performance on three out of six Combine problems; giving priority to thinking produced a pattern that matched children's on two out of six.

In contrast, without its Change schemata, the simulation could *not* solve any of the Change problems because it could not understand the transfers described in them. In other words, it had no way to map these transfers onto its conceptual knowledge concerning part–whole relations. This is because Change problems describe story situations, not simple comparisons or combining of set quantities. Without its story-situation knowledge, it could process the text base but not produce a coherent representation of the story.

Turning finally to Compare problems, the simulation was found to produce wrong operation errors on Compares 1 through 3 and on Compare 6, while correctly solving Compares 4 and 5—all using the same strategy. This strategy is one that assigns the role of SUPERSET to the set that is *least specified* in the problem. The reason it did this is rather interesting. Since it no longer understood comparison scenarios, i.e., had no *direct* mappings from comparisons to part–whole structures, it ignored the phrases containing the comparative form. The remaining parts of these lines therefore specified sets owned by no one and were hence considered specifications of the superset. For Compares 1 and 2, this translates into "Mary has 5 marbles. John has 3 marbles. How many marbles ⟨are there altogether⟩?" Under these circumstances, the simulation assigned the role of superset to the unknown quantity referenced in the last line of its representation and added the other two set quantities. Most importantly, when children performed wrong operation errors on these problems, they misrecalled the problem in just this way 77% of the time. For Compares 3 and 6, ignoring the comparative phrase in the last line translated into, e.g., "Mary has 3 marbles. ⟨There are⟩ 5 marbles ⟨altogether⟩. How many

does John have?" Here, it subtracted 3 from 5, instead of adding them as it should have. Children were also observed to make wrong operation errors on these problems, but there was no evidence of this type of misrecall in their protocols. Finally, on Compares 4 and 5, ignoring the comparative form produced a representation whose interpretation was "Mary has 8 marbles. (There are) 5 marbles (altogether). How many does John have?" In this case, the simulation assigned the role of superset to the quantity "5" and produced a negative number as its solution. Children were not observed to do this. On Compare problems, then, the simulation matched children's errors on only 4 problems. Across all 18 problems, altering its story understanding knowledge produced a match with children's performance on only 7 problems.

Deficient linguistic knowledge. To test the contribution of knowledge concerning words and word phrases to solution success, we restored the simulation's story schemata and altered instead its understanding of certain key words and phrases. The alterations were of three types: First, we altered its understanding of the word "SOME." As noted by Riley et al., among others, children often seem confused about the interpretation of this term, choosing often to completely ignore it when modeling solution performance with blocks. Accordingly, we removed the word "SOME" from the simulation's quantity word category and placed it instead in its modifier category. Essentially, this means that, to the simulation, "SOME" was no longer a word that specified an unknown quantity, but was instead an adjective. Second, as noted by DeCorte and Verschaffel (1986), children tend to ignore comparative linguistic forms when reading. This tendency was also noted in our recall data. Accordingly, the simulation was made to misinterpret "HAVE-MORE-THAN" simply as "HAVE". For example, Compare 3 was misparsed[2] as follows:

> Mary has *X* marbles.
> John has *Y* marbles.
> *X* marbles are more than *Y* marbles.
> How many marbles does John have?

Essentially this means that the comparative form is interpreted as a statement primarily about set ownership and tangentially about the relative sizes of two cardinals; nowhere is there an understanding that the original statement refers to a difference set. The third line in this misparsing was

[2] The simulation does not parse natural language, but uses a propositionalized text base as its input. In an earlier version (Dellarosa, 1986), the misparsings described here were simply used as input. In the current version, production rules produce the misparsing effect by constructing different types of proposition frames based on the expertise level of the simulation run. The examples in the text constitute transcriptions of the proposition frames constructed during a low expertise run.

not recalled by children, but it was included in our simulation to allow the MORE-THAN proposition to enter short-term memory—as it presumably does when children hear it—and hence affect the processing load. Note the difference between this treatment of the comparative and the treatment received when the story schemata are missing. Here, the comparative form is completely misunderstood as a statement of ownership; in the former case, it was ignored entirely because it was not taken as a statement of ownership nor could it be understood as a comparison scenario since no knowledge of such a scenario was present in the simulation's knowledge base.

Finally, there was some indication in our data that children have difficulty interpreting the term "ALTOGETHER," as in "Mary and John have *X* marbles altogether." There was some suggestion that children interpret this as meaning that Mary and John EACH have *X* marbles. Accordingly, the simulation's proposition processing rules were made to interpret these linguistic forms as follows: "Mary has *X* marbles and John has *X* marbles."

To summarize, in its present state, the simulation interpreted SOME as an adjective, HAVE-MORE-THAN as simply HAVE, and ALTOGETHER as EACH. With these changes in its linguistic structures, the simulation was again required to solve the 18 problems.

When the simulation's linguistic knowledge was impaired as described, it produced a response pattern that matched that of children on 15 out of 18 problems. These results indicate that the characteristic errors reported here and elsewhere in the literature primarily reflect difficulties children have in assigning interpretations to certain words and phrases used in standard word problems.

Let us begin with the Compare type problems, since these are the most straightforward cases. Recall that these problems are misinterpreted as follows:

> John has 4 marbles.
> Mary has 3 marbles.
> How many does (John, Mary) have?

In such a case, the simulation simply builds three unrelated sets, one corresponding to John's marbles, one corresponding to Mary's, and one corresponding to the set whose cardinality is requested. No superschema is built since there is no information about how these sets are logically related. As a result, none of the standard arithmetic operation rules apply, and the simulation resorts to its default rules to produce an answer. In this case, it searches memory to see if it already created a set that matches the specifications of the requested set. Finding that it does, it returns the

cardinality of the requested set, i.e., John's or Mary's marbles, as appropriate. As a result, it matched children's performance on five of the six Compare problems.

A similar situation arises when the term "ALTOGETHER" is mapped onto EACH. In this case problems that contain sentences such as "John and Mary have 12 altogether" end up being represented as follows:

John as 12 marbles.
Mary has 12 marbles.
Mary has 6 marbles.
How many does John have?

Again, the simulation ends up with four unrelated sets in memory, and no information about how they are logically related. As a result, it performs a search for a set corresponding to John's marbles, and returns "12," or the superset cardinal. Accordingly, it matched children's performance on five out of six Combine problems, the exception being Combine 5, on which our subjects committed wrong operation errors instead of given number errors.[3]

The case of Change problems is a bit more complex. In order for the simulation to solve a Change problem, it must build a coherent TRANSFER-IN or TRANSFER-OUT schema. A TRANSFER-IN schema is built if the problem describes a starting set into which objects are transferred. A TRANSFER-OUT schema is built if objects are transferred out of the starting set. A difficulty arises when the simulation does not understand "SOME" to be a quantity word. In such a case, it does not create a set when it encounters a proposition containing "Some." In Change 5 and Change 6 problems this is particularly disastrous, because "Some" describes the starting set. Without this all important set, there is not enough information to determine whether the problem describes a TRANSFER-IN situation or a TRANSFER-OUT situation. As a result the simulation again ends up with three unrelated sets (corresponding to lines 2, 3, and 4 in the problems) instead of a coherent superschema under which these sets are subsumed. In order to solve the problem, it resorts to its default rules. In this case, it can either (1) return the cardinality of the set specified in the final line of the problem (e.g., Mary's marbles) or it can (2) use the term "BEGINNING" as a cue to return the cardinal of the first set it created (i.e., the cardinal of the transfer set, line 2 of the problem). Accordingly, the simulation matched the children's performance on four of the six Change problems.

To summarize, the best match between the children's performance and

[3] It should be noted that although our subjects committed wrong operation errors on Combine 5, DeCorte, Verschaffel, and DeWinn (1985) reported that their subjects committed superset-given number errors on this problem, just as our simulation did.

the simulation's was obtained when the latter's language processing was altered, as opposed to its logico-mathematical knowledge. Two discrepancies did occur, however. The first was the fact that children were observed to make an abundance of wrong operation errors on Compare problems in addition to given number errors; our linguistically deficient model produced solely given number errors. Note, however, that the simulation did produce wrong operation errors on these problems when its story-understanding knowledge was deficient. These results suggest that children have two strategies for dealing with the difficult comparative linguistic form. The first is to simply treat it as a statement about possession (as our linguistically deficient model did) and the second is to ignore it completely (as our schema-deficient model did). In the former case, a given number error occurs; in the latter, a wrong operation error occurs. The same can be said of the term ALTOGETHER which can be interpreted either as EACH or as a command to add the numbers in the problem.

The other discrepancy occurred on Change 3 and Change 4. Here, the simulation could solve the problem (even without knowing that SOME is a quantity word). Children sometimes had difficulty with these problems, as evidenced by the given number errors observed on this problem. It is not clear why this discrepancy occurred, although it should be noted that even children do not find Changes 3 and 4 as difficult as Combines 3 through 6 or most of the Compare problems (see Figure 1).

Most important to our endeavor is the fact that the patterns of solution difficulty reported here and elsewhere in the literature could be accounted for simply by manipulating linguistic aspects of the simulation program. The major determinant of its solution characteristics was whether the nature of its linguistic processing afforded access to its conceptual knowledge. Certain wordings allow direct mapping onto part–whole structures (e.g., Changes 1 and 2); others instead require inferences about these mappings (e.g., Compares 3 through 6, Changes 5 and 6) or instruction concerning special interpretations/mappings of certain words in mathematical settings (e.g., "altogether," "some"); these problems are therefore more likely to be misunderstood.

EXPERIMENT 2

The results of Experiment 1 supported our claim that language comprehension strategies play a central role in word problem solving. In Experiment 2, we tested our claim further in two ways. First, we employed a new set of problems that better deserve to be called story problems than the impoverished versions used in Experiment 1. The problems were designed to be little vignettes, showing plausible, realistic situations and setting up a motivation for the final arithmetic question that com-

pleted the story. In addition to using structural recall as a measure of comprehension, however, we also used *question generation:* For half of the problems, children were required to generate a plausible question to end the story; for the other half, the question was presented along with the rest of the story. In order to generate a plausible question, it was necessary to understand the story, that is, understand the world as described by the story text. Question generation, therefore, should serve as converging evidence of story comprehension.

Method

Subjects. The participants were 36 second and 36 third grade children from the Jefferson County School system. The majority were white, middle class children of average intelligence. The schools were paid $5.00 for each student's participation.

Materials. Four difficult problem types from those used on Experiment 1 were chosen on which to base the stories. These included Combine 5, Change 5, Change 6, and Compare 5. These problem types were then embedded in rich story contexts. The stories were all five lines long, and their propositional content varied from 18 to 31 propositions, with a mean of 23.9 and a standard deviation of 3.9. There were 20 story problems in all, five from each of the four problem types. One of each of these were used as practice problems; the remaining 16 served as the stimulus materials. The numbers embedded in the problems were double digit numbers. Examples of the problems are presented in Table 4.

Procedure. Subjects were tested in a small, quiet room in their schools during school hours. The sessions, which lasted approximately 1 h, were recorded on tape. The child was informed of the tape recorder, but was assured that no one but the experimenter would hear the tape (i.e., parents and teachers would not). The tape recorder was unobtrusive, and no child seemed uncomfortable with its presence. Problems were presented by placing a card in front of the child on which a problem was typed and reading it out loud. The reading rate was kept slow enough to ensure that the child could follow along.

There were two experimental factors. The first was recall order: Recall before or after solving a problem. The second was the question task: Generate or listen to the final line of the problem story (Generation vs Standard question condition). These two factors were crossed to form 2 × 2 within-subject design. Each subject received one problem from each problem type under each of these conditions.

The sessions began with verbal instructions concerning the problem solving and recall tasks. The standard question condition always preceded the generation condition, and separate instructions were given prior to beginning the generation task. Following instructions, the child was given two practice problems. One problem was first solved and then recalled, and the other was first recalled and then solved. Recall was initiated by asking the subject to tell the story back. In order to ensure a reasonable amount of recall, a set of recall prompts was used whenever the child failed to respond. A subset of words from the stories was reserved for this purpose. These words consisted primarily of character names and time sequence words such as "then she." All subjects were prompted using the same words. The question generation task was initiated by asking the subject to think up a good question to complete the story. Once a question was generated, the child was asked to answer the question, that is, solve the problem.

During practice, subjects were assisted in solving the problems and clarifying the task. Once the experimental session began, no help was given other than recall prompting. The card on which the problem was typed was turned over immediately following reading. Typed on the back of the card were the numbers required to solve the problem, including the word

TABLE 4
Examples of Each Problem Type Used in Experiment 2

CB5	Jane and Susan are having a slumber party on Saturday. They will have to collect pillows for everyone before then. So far, they have collected 16 pillows altogether. Susan brought 7 from home. Jane borrowed the rest from her cousin. Q: How many pillows did Jane borrow from her cousin? Equation: 16 − 7 = ?
CH5	Jim and Michael both have toy car collections. Yesterday, they played in the attic and found 4 cars. They belonged to Jim's grandfather when he was young. He gave them to Jim as a present. Now Jim has 9 cars in his collection. Q: How many cars did Jim have in the beginning? Equation: ? + 4 = 9
CH6	Bill carries a lot of things in his pockets. He also tends to lose things all the time. Today on his way home he dropped 3 shells. When he emptied his pockets he found only 6 shells. Bill was sad because his father had given him those shells. Q: How many shells did Bill have in the beginning? Equation: ? − 3 = 6
CP5	Jane and Mimi play tennis together twice a week. They both always try hard to beat each other. Both of them decided to buy new tennis rackets. So far Jane has saved 13 dollars for her racket. She has saved 5 dollars more than Mimi. Q: How many dollars has Mimi saved? Equation: 13 − 5 = ?

"SOME." Subjects were asked to write down the numbers on their work sheet and to indicate whether they intended to add or subtract (i.e., they had to write an equation). When they solved the problem, they also indicated their answers on the sheet.

To summarize, the sessions consisted of two practice problems, four standard problems, two of which were solved first and two recalled first, another two practice problems, and four question-generation problems of which two were solved first and two were recalled first. Order of recall task presentation was counterbalanced across subjects, as was problem presentation.

RESULTS

Protocols were scored for the following: (a) solution accuracy, (b) question accuracy, both recalled and generated, (c) equation accuracy, (d) solution error type, (e) propositional recall, and (f) structural recall. The following analyses were conducted on these data. Unless otherwise stated, rejection probability was .05. One third-grader's data was lost due to a faulty tape. In all subsequent analyses, the appropriate group mean was substituted for that subject's data.

Answers. The proportion of correct answers produced by subjects as a function of question and recall condition were computed and are presented in Table 5. An analysis of variance was conducted on these data using as factors grade (Second or Third), question task (Generation and Standard), and recall condition (Solve Before Recall and Solve After Recall), with repeated measures on the latter two variables.

The analysis returned two significant results. The first was the main effect of grade, $F(1,70) = 7.07$, $MS_e = .27$, $p < .001$, indicating that third graders solved more problems correctly than did second graders. The second was the main effect of recall condition, $F(1,70) = .39$, $MS_e = .08$, $p < .05$, indicating that the subjects solved more problems correctly when solutions preceded recall than when they followed recall. The question task did not significantly influence the subjects' solution performance.

It should be noted, however, that overall solution performance of second graders under the standard condition (39%) was not appreciably different than that observed in Experiment 1 (43%). Our enriched story contexts, therefore, did not boost performance on these difficult problems, contrary to our expectations.

To test our prediction concerning the contribution of comprehension factors to solution performance, we again employed a multiple regression procedure. We constructed a regression model to predict subjects' solution performance as a function of their structural recall and question task performance. Question task performance under the standard condition meant how many questions a given subject *remembered* correctly; question task performance in the generation condition meant how many times a subject *generated* an appropriate question to complete the problem stories. Four models were required to explore the relationships among these variables: One for second graders under the standard condition, one for the second graders under the generation condition, one for third graders under the standard condition, and one for third graders under the generation condition.[4]

TABLE 5
Proportion Correct Solutions to Word Problems in Experiment 2

	Standard condition	Generation condition	Mean
Third grade			
Solve First	.67	.71	.69
Recall First	.63	.63	.63
Second grade			
Solve First	.44	.33	.39
Recall First	.33	.28	.30
Mean	.52	.49	.50

Note. Cell means are based on 72 observations (two problems from each of 36 subjects under each cell condition).

For both second and third graders, only one variable met the $p = .05$ entry level condition for predicting solution performance in the *standard* condition, that of *structural recall*, $F(1,34) = 22.45$, $MS_e = .09$, $p < .001$, and $F(1,35) = 6.51$ $MS_e = .07$, $p < .02$, respectively. Thus, a subject's ability to solve a problem depended on his or her ability to comprehend the story properly. The ability to simply remember the final line to the problem did not correlate significantly with solution performance. It should be noted, however, that while this variable accounted for 40% of the variance among third graders' solution performance, it accounted for only 16% of the variance among second graders' solution performance.

The models for the *generation* condition presented a different picture. Here *third graders'* solution performance was determined by both their ability to complete the problem story with an appropriate question ($b_q = .61$, $SE_q = .12$, $p < .0001$) and their ability to recall the problem structure properly, ($b_s = .35$, $SE_s = .16$, $p < .04$). This two-variable model accounted for 67% of the variance in solution performance. *Second graders'* solution performance, however, was significantly influenced by only one factor, the ability to complete the problem story with an appropriate question ($b_q = .33$, $SE_q = .13$, $p < .02$). Despite its statistical significance, however, it should be noted that this model accounted for only 16% of the variance in second graders' solution performance in this condition. (Moreover, the regression coefficient for structural recall was significant at the .15 level, suggesting a difficulty of statistical power.) Clearly, second graders' performance was far more idiosyncratic and variable than was third graders' performance. One interpretation is that the rather taxing task demands exceeded their processing resources.

Finally, to ensure that the relationship between structural recall and solution performance was not simply reflecting subject variability, the deviation from expected frequencies in the correct-recall/correct-answer case was computed for each subject as is described in Experiment 1. These deviations were found to be significantly greater than zero, $t(72) = 4.10$, $p < .0001$, indicating that, on the average, structural recall and solution performance correlated for each subject regardless of overall performance level. The same result was observed for the relationship

[4] Equation generation was not included in the regression models because it correlated nearly perfectly with solution performance. That is, subjects had no difficulty carrying out the computations in their equations. If they wrote an appropriate equation, they got the problem right; if they wrote the wrong equation, they got the problem wrong. The more important question is what determined the subject's ability to generate the right equation. Since equation-writing and solution performance were nearly perfectly correlated, it follows that the same variables that influenced the latter would also influence the former.

between question performance and solution performance for individual subjects, $t(72) = 6.05$, $SE = .07$, $p < .0001$.

Qualitative aspects of question generation and solution performance. The questions children generated were sensible. Only 2 out of the 288 questions children produced were "off the wall." The remainder were sensible questions, such as "How many altogether?", or "How many did Mary have left?", which sometimes did not fit the problem structure they were supposed to have comprehended.

The types of questions that children generated to complete the stories were directly related to the types of solution strategies adopted, as evidenced by the numbers in Table 6. When subjects completed a story with a correct question, they tended to produce correct answers (71%). In similar fashion, wrong operation errors tended to be preceded by wrong operation questions (71%). (A wrong operation question is one that cues the wrong arithmetic question, such as completing the subtraction Compare problem in Table 4 with "How many altogether"). Given number errors also tended to be preceded by given number questions (37%). (A given number question is one that simply requests a number stated in the problem, such as completing the Compare problem in Table 4 with "How many has Jane saved?"). This tendency is probably underestimated in our data because subjects were required to write equations to solve problems. In order to produce a given number error under these circumstances, the child had to ignore the equation he or she had written or fail to write one at all. Finally, arithmetic errors really seemed to be "accidents" in that 85% of them were accompanied by the right question.

Model Predictions of Solution Performance

As in Experiment 1, we required the computer simulation model to solve the problems used here under conditions of full knowledge, deficient conceptual knowledge, deficient story situation knowledge, and deficient linguistic knowledge. It should be noted that the problems in Experiment 2 are verbally complex, and no changes were made to the model to accommodate the complexities other than adding the necessary words to its lexicon. As a result there was some question as to whether it could handle these problems at all.

TABLE 6
Questions Generated and Solutions in Experiment 2

Question type	Response type					Total
	CO	WO	GN	AR	OTH	
CO	106	17	7	17	2	149
WO	10	35	4	0	0	49
GN	13	17	18	1	0	49
OTH	12	12	7	2	4	37
Total	141	81	36	20	6	284

Note. Frequencies are based on four observations from each of 71 subjects. CO, Correct answer; WO, wrong operation error; GN, given number error; AR, arithmetic error; OTH, unclassifiable error.

When required to solve the problems with full access to the three types of knowledge, the simulation correctly solved 13 out of the 16 problems. The three that it could not handle proved troublesome because they required verbal inferences in order to comprehend them. For example, one problem stated that seven people each purchased one book. The simulation created separate sets of "seven people" and "one book." It could not infer that in such a case, seven books would have been purchased. Another problem stated that a boy gave his sister "some pencils" and then later referred back to "the five pencils that he gave her." The simulation again could not infer that these two phrases referred to the same set. The last problem stated that two girls "traded" their doll collections. The simulation could not infer that this meant a transfer of ownership. The remaining problems were solved correctly, but with longer run times and heavier processing loads than the problems in the first experiment.

The results of the simulations under conditions of degraded knowledge are presented in Table 7. Once again, the simulation performed most like children when its linguistic knowledge was impaired (11 out of 16 possible matches). The next best match occurred when its story situation knowledge was impaired (9 out of 16); the worst match was obtained when its part–whole conceptual knowledge was degraded (6 out of 16).

The most striking thing about Table 7 is the predominance of wrong operation errors among children and simulation alike. When solving the standard word problems used in Experiment 1, children and our linguistically degraded model tended to commit given number errors. While it is not clear why children evidenced this switch in error type, examination of the simulation's protocols revealed the following: When solving the sparsely worded problems in Experiment 1, the simulation often ended up with unrelated sets in short-term memory and hence produced given number errors. Here, it seldom ended up with unrelated sets because the specification of sets were complex, inviting more comparisons to be made among them to determine which set could possibly serve as a superset. This produced a tendency toward injudicious use of the least-specified rule. Consider, for example, the Compare problem in Table 4. When trying to solve this problem, the linguistically deficient model at first ended up with three unrelated sets, just as in Experiment 1. However, these sets contained elaborate specifications. As a result, its "least-specified" rule caused it to falsely identify the last set created (How

TABLE 7
Characteristic Errors: Observed and Simulated—Experiment 2

Problem	Children's errors (most frequent)	Simulation error		
		(¯C)SL	C(¯S)L	CS¯L
Combine 5				
A	WO	WO*	SBN	SBN
B	WO	WO*	WO*	WO*
C	WO–SPN	WO*	WO*	WO*
D	WO–SPN	WO*	WO*	WO*
Change 5				
A	WO	U	—	SPN
B	WO	WO*	—	WO*
C	WO	WO*	WO*	WO*
D	WO	U	—	—
Change 6				
A	WO	U	WO*	WO*
B	WO–SPN	U	WO*	WO*
C	WO	U	—	—
D	WO–SBN	U	WO*	WO*
Compare 5				
A	WO	U	WO*	WO*
B	WO	U	—	WO*
C	WO	U	WO*	WO*
D	WO	U	U	U
Total matches (*)		6	9	11

Note. Conceptual errors constitute the most frequently observed conceptual errors based on 36 subject observations per problem. Error types: WO, wrong operation errors; SPN, superset given number errors; SBN, subset given number errors; —, no error; U, unable to derive any solution. Simulation conditions: ¯CSL, no conceptual knowledge; C¯SL, no problem situation knowledge; CS¯L, degraded linguistic knowledge concerning key words and phrases.

* Problems for which the simulation matched children's error patterns.

many . . .) as a superset. Recall that this rule compares the specifications of a group of sets and identifies the one that shares many features with other sets but is *least* specified as the superset. In our Compare problem example, the specifications of the last set (i.e., dollars saved by Mimi) overlaps with the other two (i.e., dollars saved by Jane for her racket; dollars saved by Jane that are more than Mimi's), but it contains fewer specifications. Moreover, the latter two sets are disjoint. Therefore, the simulation erroneously identified the dollars owned by Mimi to be the superset whose cardinal had to be computed. As a result, it added the two numbers.

To summarize, with intact knowledge, the simulation solved all the problems that did not require world knowledge or sophisticated verbal inferences. The best match with children's error patterns was obtained when its knowledge concerning key words and phrases were impaired. More importantly, when its verbal knowledge was impaired but the problem texts were rich, its problem solving strategies changed. Rather than simply terminating the problem solving episode with unrelated sets, it attempted to use information from the problems' rich text bases in order to build part–whole schemas. It is possible that children attempted the same strategies, causing a switch from given number to wrong operation errors in their protocols.

DISCUSSION

Why are word problems so difficult? The results reported here suggest that text comprehension factors figure heavily in word problem difficulty. In Experiment 1, solution performance was found to mirror structural recall, indicating that solution strategies are dictated by the quality of comprehension achieved. Comprehension, in turn, appeared to be influenced by the nature of the language used in the problem text. Problem texts that contain certain linguistic forms are particularly difficult for children to solve. These include forms such as "SOME," "How many more *X*'s than *Y*'s?", and certain uses of "altogether." Our structural recall results show that these forms are often misinterpreted by children and, moreover, misinterpreted in certain ways. Our simulation results suggest that common solution error patterns are directly related to the linguistic sophistication possessed by the solver. The empirical and simulation results of Experiment 2 indicated that solution strategies can be directly influenced by the nature of the problem text.

More importantly, the simulation results of the two experiments clearly show the interaction of text characteristics and knowledge in determining solution strategies. Robust linguistic knowledge produced successful solution attempts regardless of text characteristics, but the attempts were more time and resource-demanding. Poor linguistic knowledge, however, produced poor solution performance, but the nature of the errors committed depended on text characteristics. Sparse texts were associated with given number errors; rich texts were associated with wrong operation errors and a preference for a more global super-set/subset conceptualization of the problem situations. More important, the primary benefit of proper linguistic knowledge in this domain appears to be in facilitating access to conceptual knowledge concerning part–whole relations. Such access allows the problem solver to construct large, cohesive problem representations in which the relations among individual sets are clearly specified.

This description of successful arithmetic problem solving as the building of large, cohesive structures is consistent with descriptions of problem

solving in other "adult" domains, such as physics (e.g., Chi, Feltovich, & Glaser, 1981) and chess (Newell & Simon, 1972). In these domains, expertise development is typically considered a matter of constructing problem type schemata and using such structures to guide comprehension. Indeed, the models of arithmetic problem solving expertise proposed by Briars and Larkin and Riley et al. are based on this very notion. From this view, then, the reason "How many more *X*'s are there than *Y*'s" is difficult for children because it depends on the possession of a part-whole superschema in memory to guide comprehension. This view, however, does not explain adequately why a simple change in wording, such as that used in Compare 4 as opposed to Compare 5, should improve performance dramatically, as do other language manipulations (e.g., Hudson, 1983; DeCorte et al., 1985). Instead, it appears that certain verbal formats allow contact to be made with superschema knowledge while others do not. Results such as these suggest that word problem difficulties may be akin to reasoning fallacies. For example, dramatic differences in reasoning about the logical conditional (if p, then q) have been produced by manipulating problem format (Cheng & Holyoak, 1985; Cheng, Holyoak, Nisbett, & Oliver, 1986; Johnson-Laird & Wason, 1977; Reich & Ruth, 1982). Typically, adults exhibit better performance when reasoning tasks are presented in concrete formats (e.g., envelopes and postage requirements) rather than abstract, formal formats (e.g., letters and numbers). Moreover, the strategies adults employ change as a function of the stimulus format. For example, Reich and Ruth (1982) reported that adults tend to match the terms mentioned in abstractly stated rules, but to verify concretely stated rules—the latter being the more appropriate strategy for the task. The tendency to produce given number errors in Experiment 1 may also be instances of this "matching" strategy: When faced with problems whose text bases are sparse and incomprehensible due to ambiguous linguistic terms, children may choose to ignore those terms and, having no other useful information available, simply match actor's names with numbers stated in the problem. (Our simulation was found to employ this strategy under these conditions.) If this were the case, then children could be described as using the same strategies adults use to solve problems under conditions of uncertainty in unfamiliar domains.

Linguistic and resource factors, however, are not the whole story. How well children perform on word problems depends, of course, on their formal knowledge of the rules and operations in the problem domain and on their level of conceptual development. These factors have been extensively studied both in the developmental and problem solving literature (Piaget, 1970; Greeno, 1978). The well-demonstrated significance of such format factors, however, need not obscure the role played by comprehension factors. The experiments reported here provide rich evidence for their importance and help us derive a fuller understanding of the processes involved in problem solving and reasoning. While Experiment 1 pointed to the contribution of linguistic factors, Experiment 2 demonstrated the importance of situational understanding to solution success. If a child understood the problem story well enough to generate an appropriate final question, he or she was very likely to derive the correct solution to the problem. Investigations into the role of situational understanding are currently underway. Reusser (1985) is developing a model in which a representation of the story situation is constructed prior to construction of the arithmetic problem representation. Kintsch (1988) has tried a different approach wherein arithmetic rules as well as general situational constraints are integrated from the very beginning of the problem solving attempt. The details of these approaches, however, are beyond the scope of the present report.

The implications of the present paper go far beyond the domain of children's arithmetic. The various determinants of problem difficulty that were explored here are likely to be important in any domain where problems are described in verbal formats. On the basis of the present findings, it would seem worthwhile to investigate the role of text comprehension factors in such domains. Moreover, the theory of arithmetic word problem solving investigated here may serve as a model for theorizing about problem solving.

REFERENCES

Briars, D. J., & Larkin, J. H. (1984). An integrated model of skill in solving elementary word problems. *Cognition and Instruction*, 1, 245–296.

Carpenter, T. P., Corbitt, M. K., Kepner, H. S., Lindquist, M. M., & Reys, R. E. (1980). Solving verbal problems: Results and implications for National Assessment. *Arithmetic Teacher*, 28, 8–12.

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

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

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

Dellarosa, D. (1986). A computer simulation of children's arithmetic word problem solving. *Behavior Research Methods, Instruments, and Computers*, 18, 147–154.

DeCorte, E., & Verschaffel, L. (1986). *Eye-movement data as access to solution processes of elementary addition and subtraction problems*. Paper presented at the meetings of the American Educational Research Association, San Francisco, April.

DeCorte, E., Verschaffel, L., & DeWinn, L. (1985). The influence of rewording verbal problems on children's problem representation and solutions. *Journal of Educational Psychology*, 77, 460–470.

Fletcher, C. R. (1985). Understanding and solving arithmetic word problems: A computer simulation. *Behavior Research Methods, Instruments, and Computers*, 17, 565–571.

Greeno, K. G. (1978). Nature of problem-solving abilities. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (Vol. 5, pp. 239–270). Hillsdale, NJ: Erlbaum.

Hudson, T. (1983). Correspondences and numerical differences between disjoint sets. *Child Development*, 54, 84–90.

Johnson-Laird, P. N., & Wason, P. C. (1977). *Thinking*. Cambridge: Cambridge Univ. Press.

Keppel, G. (1973). *Design and analysis*. Englewood Cliffs, NJ: Prentice–Hall.

Kintsch, W. (1988). The role of knowledge in discourse comprehension. *Psychological Review*, in press.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, 92, 109–129.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice–Hall.

Piaget, J. (1970). *Genetic epistemology*. New York: Norton.

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

Reusser, K. (1985). *From situation to equations: On formulating, understanding, and solving situation problems*. Technical Report No. 143, Institute of Cognitive Science, University of Colorado, Boulder.

Riley, M. S. (1981). *Conceptual and procedural knowledge in development*. Unpublished Masters thesis, University of Pittsburgh.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem solving ability in arithmetic. In H. P. Ginsberg (Ed.), *The development of mathematical thinking*. New York: Academic Press.

Stigler, J. W., Fuson, K. C., Ham, M., & Kim, M. S. (1986). An analysis of addition and subtraction word problems in U.S. and Soviet elementary mathematics textbooks. *Cognition and Instruction*, 3, 153–172.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

(Accepted January 21, 1988)