ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

PRODUCTION NOTE

University of Illinois at
Urbana-Champaign Library
Technical Report No. 166

THE EFFECTS OF INFERENCE TRAINING AND PRACTICE ON YOUNG CHILDREN'S COMPREHENSION

Jane Hansen
University of New Hampshire

P. David Pearson
University of Illinois at Urbana-Champaign

April 1980

Center for the Study of Reading

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN
51 Gerty Drive
Champaign, Illinois 61820

BOLT BERANEK AND NEWMAN INC.
50 Moulton Street
Cambridge, Massachusetts 02138
Inference Training
The Effects of Inference Training and Practice on Young Children's Comprehension

Several years ago, Guszak (1972) reported a study in which he found that children were best at answering the kinds of questions that teachers asked most often; teachers tended to ask rather direct "literal" comprehension questions about four times as often as inferential or interpretive questions; students in Guszak's study performed much better on literal than inferential or interpretive questions. The results of that study have always intrigued us because it is not clear whether the students' superior performance on literal comprehension probes was due to the fact that such questions are inherently easier to deal with than are inferential questions or, alternatively, because students simply get much greater opportunity to practice answering such questions. We have always wondered whether it would be possible via instruction to alter children's facility to deal with inferential comprehension probes. Further, we wondered whether practice in answering a steady diet of inferential questions would be sufficient alteration, or would such practice result in only a surface mindset that would not transfer to other situations? Maybe an alternative approach would be to provide training--in contrast to simple practice--in the process of inferencing.

Until recently, however, we were reluctant to undertake such an experimental endeavor, largely because we had no well-motivated way of explaining any conceivable set of results we might obtain and because we had few, if any, operational guidelines for distinguishing between literal and inferential question probes. Recent developments in the study of cognitive processes involved in text comprehension and inferencing and in
the technology of question probe development have provided the theoretical framework into which such a study might be set.

Consequently, we set out to investigate whether or not direct intervention strategies in the form of alterations in the questioning environment in which students spend their instructional time would result in systematic changes in their ability to establish inferential relationships.

Our framework is derived metaphorically if not directly from notions that have been developed recently to explain comprehension and memory processes (e.g., Schank, 1972; Minsky, 1975; Rumelhart & Ortony, 1977; and Anderson, 1977). What has been so appealing within these "schema theories" are the procedures that they hypothesize for explaining how new information (e.g., that which might come from a text) is meshed with existing knowledge (i.e., those knowledge structures that comprise a reader's long term memory) in the process of comprehension.

Of particular interest is the explanation of inference within schema theory. The most common kind of inference, slot filling, works like this.\footnote{In order for an idea in a text to be understood, it has to instantiate a schema (a general knowledge structure) in the reader's long term memory. For example, sentence (1) might instantiate a building schema (along with a carpenter schema and some affective schemata like pain, etc.).}

(1) The carpenter became angry when he hit his thumb instead of the nail.

Texts are never completely specific in reporting an incident (we admit that our example is particularly sketchy). Authors seldom report what they
think readers already know. Notice that in (1) the instrument for the hitting is unspecified. Yet most adults and children would probably respond "hammer" if asked question (2).

(2) What did the carpenter hit his thumb with?
This would occur even though the instrument is unspecified in the text precisely because when the building schema was instantiated (by cues like carpenter and nail), the default assignment for the instrument slot within the building schema was hammer. Hence, in the act of schema instantiation, a lot of excess baggage (in the form of default assignments of values to slots not specified by the text) gets carried along and brought into focus or readiness for further processing. That this is true can be demonstrated by the puzzlement we would invoke if sentence (1) were followed by sentence (3) in the text,

(3) "I've just got to get a new rock," he murmured to himself, or our disbelief if (1) were followed by (4),

(4) "I've just got to get a new saw," he murmured to himself.

Default assignments to variable slots not specified in a text represent our best guesses about what should fit with the schema we have instantiated or brought into focus. The practice of assigning default values is ubiquitous. We can hardly process a sentence of text without doing so. Consider what happens to you when you read sentence (5), from Rumelhart (in press).

(5) Business had been slow since the oil crisis.
You can hardly resist the temptation to fill the general business schema with a particular type of business, such as automobile, fuel, recreation, etc. But notice that other cues in the text, particularly the values that fill other variable slots, influence the particular default value that we will assign to an unspecified variable slot. So if sentence (6) follows sentence (5), we fill our business slot with a value different from that we would use if sentence (7) follows sentence (5).

(6) Nobody seemed to want anything elegant anymore.

(7) Nobody seemed to want to travel very far from home anymore.

(6) predisposes people to fancy cars; (7), to the recreation industry. The point is that default assignment is not made independently of the text. In fact, knowledge in our long term memory interacts with the information in the text to tune these assignments; often we find that as we read further, we must alter our original assignments to resolve a contradiction with some new information.

With respect to adults, we might respond to this account of inference with an acknowledgement of its plausibility and some suspicion about its importance. After all, on what other basis can adults make inferences save by reference to their existing knowledge? But with children the matter is not so simple. First, we would admit that children's more limited store of prior knowledge would make such slot filling inferences less probable. Second, we might question either their ability or inclination to draw such inferences at all. In other words, can they? And if they can, do they do so spontaneously?
That adults can and do draw such inferences in the process of comprehension is well supported in the literature. Kintsch (1974) developed two versions of a passage, one complete and the other missing a key item of information. Comprehension was checked 20 minutes after the reading. Latency for answering questions requiring the missing item of information did not differ across versions. Kintsch concluded that the textual information had not been stored in memory in intact textual form. Instead it was integrated with existing knowledge structures during the encoding process; thus, the implied information was as "ready" as the explicit information when comprehension was checked. Frederiksen (1975) gave adults different sets of directions prior to reading passages. He found that subsequent written recalls varied as a function of prereading directions. Both of these studies have been interpreted as supporting a constructive view of memory; i.e., that interpretation of incoming data occurs at the point when it is encoded into memory. However, the issue of when inferences are made, at encoding or retrieval, is not settled. Certainly the work of Spiro (1977) suggests that many inferences can be drawn at retrieval; and the work of Bartlett (1932) as well as the Frederiksen study cited earlier (1975) suggest that memory for a text is characterized by more intrusions from prior knowledge as the time interval between reading and recall increase. While there may be some debate over when inferences are made (at encoding or retrieval), the common element in all these studies is that inferences are an inevitable part of the comprehension process.
If inferencing is a necessary process involved in comprehension, then children must possess this skill. Studies about children's inferencing skills have suggested that they do differ from adults'. Many educators have believed for some time that the difference is qualitative; children are not capable of drawing the same types of inferences as adults (Piaget, Inhelder, & Szeminsko, 1960; Bloom, 1956). However, this contention is currently viewed with some degree of doubt. The difference may be quantitative. Two main avenues of research have led to this tentative conclusion. However, the two lines of research draw alternative conclusions regarding why children make fewer inferences than adults.

One avenue of research has been pursued by Trabasso and his colleagues. Trabasso, Nicholas, Omanson, and Johnson (Note 1) proposed a taxonomy of inferences utilized by children in comprehending stories. One of their goals was to discover whether the development of inferential ability is a function of the content of a child's schemata. Omanson, Warren, and Trabasso (Note 2) found the effect of prior knowledge to be prominent. Two groups of children, ages five and eight, were tested, and it was found that when equivalent levels of veridical recall of text occurred, the eight-year-old children did draw more inferences than the five-year-olds. Since inferential limits were not due to memory limits, Omanson et al. hypothesized that they were a consequence of insufficient prior knowledge. In other words, young children do not lack the ability to draw inferences, nor do they lack memory capacity. What they often lack is prior knowledge which, in turn, may limit their ability to draw an inference in a particular situation.
Another series of investigations has been conducted by Paris and his associates. Paris and Lindauer (1976) presented seven-year-old children with sentences such as (8) and (9).

(8) Our neighbor unlocked the door. (implicit instrument version)
(9) Our neighbor unlocked the door with a key. (explicit instrument version)

Half of the children received the implicit instrument version and half received the explicit version. Testing consisted of presenting either explicit or implicit cues and asking the children to recall the sentences. The children who had been given the implicit instrument versions were not able to recall the sentences when the cues (e.g., key) were provided. A follow-up study was a successful attempt at "rigging" the setting so that the children acted out implied relationships. For example, upon being presented the sentence in the implicit instrument version, the student was asked to dramatize the action. However, the word key (which would be the cue used later for testing) was not mentioned. Following the dramatizations, these children understood the implicit versions as well as others understood the explicit. The authors concluded that young children do not spontaneously build semantic relationships even though they are capable of doing so. These children did possess the necessary background knowledge to draw the required inferences, but they did not spontaneously integrate the new information with the old.

Thus, two possible explanations have been offered to explain the quantitative differences between adult and child inferential performance—extent of prior knowledge and spontaneity in drawing inferences.
Brown (1977) believes that the skills which are not used spontaneously need to be induced in children. In order for them to use their knowledge, they must employ processes of prediction, planning, checking, and monitoring. According to Brown, research needs to focus on the development of programs that will train children to apply these processes. It seems that children spontaneously draw inferences in their daily activities; they begin analyzing similarities and differences when they are infants. However, they do not make these inferences as consistently when confronted with reading tasks.

The two intervention methods tested in the present study emanated from Brown's suggestion that we need to create within children the realization that they must and can draw inferences between print and prior knowledge.

One method in the present study, the Strategy Method, was specifically intended to capitalize on the importance of prior knowledge and to induce spontaneity by increasing in the children an awareness that they can make inferences between their previous experiences and the stories they read. The method utilized is consistent with Pearson and Johnson's (1978) recommendation that predictions prior to reading help to highlight the students' related knowledge, thus increasing the likelihood that while reading they will consciously try to integrate text and prior knowledge.

Two recent intervention studies were intended to systematically capitalize on students' extralinguistic knowledge as a means of improving comprehension. Prior to reading passages, Swaby (1977) presented sixth grade children with a written statement designed to create a connecting
link between their possible previous experiences and the text. The procedure did not facilitate comprehension nor, specifically, inferential comprehension. Swaby did find, however, that a prereading treatment of helping students understand key concepts in the passage facilitated the post-reading comprehension of low ability students. Schachter (1978) added the dimension of discussion to the prereading activities. His fifth grade students discussed questions which used the word *you* as the means of creating connecting links between previous experiences and the main ideas of the upcoming stories. He found that the procedure enhanced inferential but not literal comprehension on questions from the instructional stories. However, standardized test scores were not affected.

Thus, only limited support has been found for improving comprehension by employing prereading activities which focus on relating prior knowledge to text.

The other instructional method utilized in the present study, the Question Method, was intended to induce spontaneity of inferencing by providing considerable practice in answering inferential questions. Guszak (1972) found that reading teachers, when asking discussion questions, asked inferential questions only about 15% of the time. One explanation for the poor performance of children on inferential questions in tests (National Assessment of Educational Progress, 1976) may be that they receive insufficient practice in such tasks. If considerable practice is provided, the ability to draw inferences may improve. Also, if children anticipate questions which require integration, they may develop a mindset toward print which focuses on interpretation rather than remembering facts. Such an approach also seems
reasonable in the light of much recent research suggesting that the single most potent factor in the instructional environment is engaged time on task (Rosenshine, 1978).

Considerable research has been conducted on the effectiveness of various levels of questions upon comprehension. Bloom (1956) delineated a hierarchy of question types which has been utilized and modified for nearly 25 years. Many correlational studies have attempted to establish a relationship between some of these levels of questions and achievement. Medley (1977) reviewed teacher-effectiveness research and concluded that low SES students in grades K-2 benefit from low-level questions, but that no pattern emerged regarding high SES children.

Winne (1979) reviewed 17 experimental studies concerned with the effects of higher cognitive versus factual questions on student achievement and concluded that question type makes little difference to student achievement. This was considered to be significant because there seems to be much professional consensus that teacher questions have a major impact on achievement. Andre (1979) concluded that only when the questions lead the learner to process the material in ways she/he would not otherwise have done will such questions influence learning.

Thus, it is questionable whether a method which involves the use of higher-level questioning can succeed. However, if the intent is to improve inferential ability, then practice in establishing relations should be more productive than practice in remembering facts.

Although the findings from various research studies regarding children's inferencing ability is mixed, certain predictions seem plausible vis-à-vis
the present experiment. Extrapolating from the work of Brown, Trabasso and his colleagues, and Paris and his associates, we reasoned that a pre-reading strategy that focused on making connections between prior knowledge and information in print would heighten children's awareness of inferencing possibilities, thus increasing the likelihood that they would draw inferences spontaneously. Second, arguing from the time on task position, we thought that increasing the number of opportunities that students had to draw inferences might also increase their spontaneous inferential behavior. Third, we thought that this focus on inferencing present in both treatments would result in superior processing of information explicitly stated in text. This, we reasoned, would occur because most inferences require an integration between prior knowledge and text rather than an exclusive emphasis on prior knowledge. Hence the manipulation of textual information required in inferencing should result in better comprehension of that information itself. Fourth, we anticipated that while we could alter children's spontaneous inferencing behavior, we would not be able to eradicate the prior knowledge effect (cf. Omanson et al., Note 2). Hence post treatment transfer effects would not be as strong in unfamiliar as they would be in familiar material. Fifth, we anticipated that this same growth in inferential behavior, relative to a control group receiving a diet of literal probes, might not be sufficient to wipe out the difference researchers have found between performance on literal and inferential comprehension probes.
Method

Subjects

The subjects were 24 second grade students attending one elementary school in a middle class suburb of St. Paul, Minnesota. Two criteria were established prior to selecting the students: (a) the children were all to be reading at the same instructional level; and (b) they were all to be reading at, or slightly above, grade level. Teacher judgment and individual assessments administered by the school's reading supervisor determined which 24 children from a pool of 90 second grade children would participate. The students were divided into three instructional groups of eight children each. The experimental condition (Strategy, Question, or Control) was randomly assigned to each intact group.

Instructional Procedure

As indicated there were three treatment groups. The Strategy group differed from the other two primarily in the focus on integrating text and prior knowledge information prior to reading. The Question group's primary focus was a steady diet of inferential questions. The Control group received a traditional mix of literal to inferential probes (about 4:1).

Each group met daily with one of the experimenters as teacher. A total of ten stories was taught (each over a period of four days), with all groups receiving identical vocabulary introduction on Day One (See Table 1).

Insert Table 1 about here.

Day Two consisted of an introduction to the story. For the Question and Control groups, suggestions in the teacher's manual were followed. For
the Strategy Group Day Two was the day during which its unique treatment was administered. The strategy was based upon a weaving analogy: weaving new information into old information already existing in one's brain. In order to make the analogy graphic for the young children, each child received two concrete devices: (a) a piece of gray paper (his/her brain) which was slit into three strips and (b) three strips of brightly colored paper to represent new knowledge. The experimenter selected three main ideas from the story and introduced them one at a time via a two-step procedure: First, a question was asked which related to some possible previous experiences of the children and second, the children were required to hypothesize something similar that might happen in the story. After oral responses to the first question, each child recorded his/her own experience on the first strip of his/her brain. Then hypotheses pertaining to the story were discussed and each child wrote his/her hypothesis on one of the colored strips. Also, at each session the strategy of relating the new to the known was reviewed as an abstract concept. As a follow-up activity the children wove the colored strips into their "brain."

Day Three consisted of the guided reading of the stories with the focus being on the difference between the questions asked of the groups. The questions asked of the Strategy and Control groups were formulated by: (a) tabulating the suggested questions in the teachers' manual to derive the percent of literal/inferential probes, and (b) if the questions did not correspond to the prevalent (Guszak, 1972) ratio of approximately 4:1, questions were altered to achieve that proportion. Minor changes were made for approximately half of the stories. The Question group received its
experimental treatment during this guided reading. All of the questions requiring inferences that were asked of the other groups were used, plus all of the literal questions used by the other groups were rewritten to require inferences. After reading the story, all students completed seatwork in the form of a ten-question worksheet (cf. page 17).

Each of the three types of training questions is illustrated with respect to a segment of a passage used in the training. Example (10) is the passage segment (Clymer & Vilscek, 1969, p. 35). (11) is a literal item used by the Control group, (12) is an inferential item used by the Question group, and (13) represents an item used by the Strategy group.

(10) Three men came to get their hair cut, but Stanley barked at them. The barber looked at William. "Boy," he said. "Isn't that your dog?"

"No," he said. "He just followed me. He lives next door."

"Well," the barber said, "that dog is keeping people out of my shop. There are people here ahead of you, but I'll cut your hair now." . . .

"It (your haircut) looks fine," Father said. "You weren't gone long. That's what a boy can do when he decides to hurry."

"It's the barber who hurried most of all," said William.

(11) What did Stanley do when three men came to get their hair cut?

(12) Why did Stanley bark at the three men?

(13) Think of something you did that you didn't tell your parents. What didn't William tell his father?

Day Four consisted of the skill/phonic activities that were suggested in the manual and which followed the curriculum of the school.
Instructional Materials

All groups used the materials normally included in the district program, the Ginn 360 basal reading program. Workbook pages, worksheets, Magic Circle books, and games that were correlated with the stories were used with all groups in order to make certain that the only differences among the three groups involved the unique treatments administered to each and to maximize the ecological validity of the setting in which the treatments were administered.

Testing Materials

Experimenter designed pretest measures. Prior to the experiment, each student read two stories and answered 20 comprehension questions following each story. There were two kinds of questions: literal and inferential. The procedures for developing and scoring these questions were the same as those used for the experimenter designed posttest, and are described below in the section dealing with that posttest (cf. pp. 17-19).

The pretest was given to further corroborate the equivalence of the groups on behavior germane to the treatments. When an ANOVA was conducted using a correct/incorrect criterion for scoring responses, no significant differences among the groups were found, $F(2,21) = 1.986, p > .10$; even so, we were concerned about possible advantages for the Question group ($M = 20.50$) over the Strategy ($M = 16.38$) and Control ($M = 16.38$) groups. In addition, an ANOVA performed using a second scoring scheme which weighted responses to inference questions on a 4 point scale (cf. p. 19) indicated that there were important and statistically significant differences among the treatment groups, $F(2,21) = 3.95, p < .05$, $M_{	ext{Quest}} = 44.75$, $M_{	ext{Strat}} = 34.25$, $M_{	ext{Cont}} = 34.25$. 


M Cont = 38.38, and M Strat = 34.50. In other words, one of our proposed experimental treatment groups was particularly adept at inferential behavior; the other, particularly inept. Consequently analyses of posttest measures were run using both ANOVA and analysis of covariance (ANCOVA) techniques, with pretest scores and weighted pretest scores as covariates for posttest measures.

Comprehension questions following the instructional stories. Experimenter-designed worksheets, which consisted of ten questions per story, were analyzed for the last five instructional stories. The first five stories were not analyzed on the grounds that any treatment effects would not have had time to take effect. Each set of ten questions consisted of two explicit questions, two inferential, two strategy, and four parallel to each treatment. The explicit, inferential, and strategy questions were constant across groups but had not been asked during any previous discussions. The four parallel questions were different for each group and were repetitions of questions asked in previous discussions with that group. They were intended to reinforce the instructional treatments. The following directions were given to the children:

These questions are about the _______ story. Some of these questions are identical to questions that our group has already answered and some are different. Please read the questions and carefully write the answers. If you need help, please ask me. Mrs. _______ (the teacher supervising the independent work) will not be able to help you with these sheets. Please do your very best.

All responses were scored as either correct or incorrect. An inter-judge percentage of agreement of 92 was attained for the scoring.
Experimenter-designed posttest. An experimenter-designed test was administered individually to all the children. Each child read silently two expository selections which differed in terms of the familiarity of the topic, one being more familiar and the other less. The original familiarity differences were made intuitively but these differences were confirmed by administering an eight-question prior-knowledge test two weeks prior to the reading of the selections. Significant test-scores differences favoring the familiar topic were found on the prior-knowledge tests.

The stories were selected from end-of-first grade basal readers not used in the participating school. The revised Spache formula was applied to all the stories, and they were judged to be of reasonably equal difficulty. The stories were altered when necessary in order to achieve equal length (approximately 250 words). A primary typewriter was used to type them in similar formats of 2½ pages each.

The children read each story in a separate session after being given the following oral directions:

I have a story about ________ for you to read to yourself. Please read it carefully and don't hurry. If you don't know some of the words, you may ask me. When you are finished, return it to me and I will ask you some questions. Please do your very best.

The experimenter recorded all responses to 20 oral, open-ended comprehension questions following each story. Ten questions could be answered by reiterating information explicitly stated in the text (literal) and ten required inference to prior knowledge. The two types of probes were interspersed to reflect the sequence of the story. The literal questions were
generated according to Bormuth's (1969) procedure for generating wh-
transformations of story statements.

Inference questions were generated using the following paradigm:
(1) Text segments were identified for which we felt the text provided no
useful explanation of the phenomenon in question. (2) Based upon our own
knowledge of the topic, we generated what we felt was a useful explanation.
(3) We asked three judges to rate the usefulness of these augmentations
in understanding the selection (their average rating on a scale from 0 to
5 was 4.0 for all selections). (4) We then used Bormuth's procedure to
generate a question which tapped the relation between the statement actually
in the text and our augmentation. It should be added that students never
actually saw the augmentations; they were used only to insure comparability
of question generation procedures across literal and inferential items.

These two item types are illustrated with respect to text (14), an
actual portion from one of the posttests. The italicized portions represent
the augmentations added to the text in order to create inference items.

Question (15) is a literal item; (16), an inference item.

(14) Many dogs learn to work for man. A good work dog is not hard
to train. But a trainer must work with him from the time he is
just a pup because if he waits until the dog is older, the dog
will have already developed other habits which would have to be
unlearned.

(15) How hard is it to train a good work dog?

(16) Why must a trainer work with a dog from the time he is just a pup?

The answers were scored using two different sets of criteria. For
the first analysis all answers were coded as being either correct or incorrect.
Inference Training

A reliability check yielded an inter-scorer percentage of agreement of 90.
The split-half reliability of the total test was .743.

For the second analysis, the following five-point scale was used to score the inference questions:

4 - A correct answer: The answer is a reproduction of, or is synonymous to, the inserted inference statement.

3 - A correct answer: The answer is based upon the inserted inference statement but is somewhat broad, specific, or incomplete. It relies too heavily on either text or prior knowledge, rather than a balanced integration of the two.

2 - An incorrect answer: The answer is related to the inserted inference statement but totally omits reference to either script or text; i.e., no inference was drawn.

1 - An incorrect answer: Such as copying from other parts of the text or a "wild guess."

0 - No response.

On the categorization of responses, the obtained inter-scorer percentage of agreement was 89.7.

Free-recall measure. A free-recall measure was administered individually at the conclusion of the study. Each student read silently an end-of-grade-one basal reader expository selection. These instructions were given orally:

I have a story about food for you to read to yourself. If you don't know some of the words, you may ask me. When you are finished, return it to me and I will ask you to tell me everything that you can remember about the story. Also, tell me anything that the story made you think of. Please do your very best.

The students' recall protocols of the story were taped and later transcribed.
Before the analysis could be done, a template proposition base was established for the text (Turner & Greene, Note 3). Then scoring was accomplished using a "form of recall" analysis in order to assess the prevalence of inferences. The forms of recall were: (a) textual information (facts repeated from the text and facts attained by combining parts of the text), (b) scriptal (knowledge-based) information (scriptal-textual inferences and scriptal elaborations), and (c) intrusions (any erroneous statements). Each student's version was loosely analyzed rather than propositionally segmented. Regarding the categorization of recall segments, the obtained agreement between two independent judges was 92%.

Reading Test of the Stanford Achievement Test. The vocabulary and reading comprehension subtests of the Stanford Achievement Test, Primary Level 1, Form A, were administered to the 24 students as a large group at the completion of the experiment.

Results
For most, but not all, of the posttest measures three separate analyses were conducted: (a) a straightforward analysis of variance (hereafter, ANOVA), (b) an analysis of covariance using correct/incorrect scores from the experimenter designed pretest (hereafter, ANCOVA), and (c) an analysis of covariance using weighted implicit scores from the pretest (hereafter, weighted ANCOVA).

Comprehension Questions Following Instructional Stories
The unadjusted and adjusted means for these analyses are reported in Table 2. For this dependent measure, separate analyses were conducted for
Inference Training

each of the three question types using scores on the two common questions of each type summed across the last five instructional stories.

Insert Table 2 about here.

The ANOVA revealed a significant effect for treatment on the inference questions, $F(2,21) = 12.149$, $p < .01$, with the following means: $M_{Strat} = 7.00$, $M_{Quest} = 5.75$, $M_{Cont} = 3.50$ (see Table 2). The Newman-Keuls post hoc procedure computed at the .05 level of significance revealed no difference between the Strategy and Question groups, but each experimental group did exceed the Control. The ANCOVA confirmed the ANOVA results, but contrasts following the weighted ANCOVA revealed an additional significant difference between the Strategy and the Question group favoring the Strategy group.

On literal questions, the ANOVA revealed a treatment effect, $F(2,21) = 5.523$, $p < .05$, suggesting that the two experimental groups performed better than the control group. The Newman-Keuls test confirmed this suggestion: The two experimental groups performed at comparable levels and each surpassed the Control. Both ANCOVAs revealed the same patterns of results. On strategy questions, the ANOVA indicated a Strategy > Question > Control ranking, $F(2,21) = 4.365$, $p < .05$. The results from both ANCOVAs were similar.

In summary, on measures related to the stories used in the experimental treatments, both intervention techniques elicited superior comprehension when compared with the Control group, even on the literal measure, which might be thought to favor the treatment given to the Control group. In addition, where differences existed between the two experimental treatments, they tended to favor the Strategy group.
Experimenter-Designed Posttests

Two different sets of scores were created from the experimenter-designed posttests: correct/incorrect and weighted implicit scores. For the correct/incorrect scores, ANOVA and ANCOVA procedures were employed; for weighted implicit scores, ANOVA and weighted ANCOVA procedures.²

Correct/incorrect scores. For this set of scores, the analysis had three factors: level of familiarity and type of comprehension question were within-subject factors and treatment was a between-subjects factor. The ANOVA revealed no interaction effects. Significant main effects were found for familiarity, $F(1,21) = 103.667$, $p < .001$, and question type, $F(1,21) = 5.579$, $p < .05$, but not for treatment, $F(2,21) = 1.535$, $p > .05$. The unadjusted means, reported in Table 3, revealed predictable effects for familiarity and question type: familiar stories elicited more correct responses than unfamiliar; literal questions more than inferential. With respect to treatment, these transfer items did not yield the same pattern

Insert Table 3 about here.

Weighted-implicit scores. Using the measure more sensitive to inferential processing, we decided to omit the overall analysis to examine scores separately within levels of topic familiarity. This decision was made a priori on the grounds that any differences due to treatments would be more
likely to surface in response to items accompanying the familiar topic, wherein, presumably, prior knowledge would be more likely to influence results. The analysis supported this assumption. According to the ANOVA, there were no differences among the three treatments on weighted implicit scores for the unfamiliar topic, $F(2,21) = 1.514, p > .05$. The weighted ANCOVA results yielded the same conclusion, $F(2,20) = .307, p > .05$ (see Table 4 for unadjusted and adjusted means). However, on the familiar topic, both the ANOVA, $F(2,21) = 5.227, p < .05$, and the weighted ANCOVA, $F(2,20) = 4.157, p < .05$, revealed a treatment effect. Post hoc tests (Newman-Keuls) on the weighted ANCOVA revealed that the Question group was superior to the Control group, $p < .05$, but not to the Strategy group. The Strategy versus Control post hoc comparison was marginally significant, $p = .066$.

Unlike the data from the instructional story items, few comparisons on these transfer items revealed an advantage for the experimental treatments. The comparison that did was embedded in contexts (weighted-implicit scoring and a familiar story) in which everything was stacked in favor of the experimental treatments, indicating a generally weak transfer effect.

**Free-Recall**

On the free-recall data one-way ANOVAs and both ANCOVAs were computed on each of the three forms of recall with treatment groups serving as the single independent variable. The ANOVA indicated that there were no significant effects among the protocols of the three treatments on any of the recall categories: not for textual information, $F(2,21) = .212, p > .05$;
Inference Training

scriptal information, $F(2,21) = 0.423$, $p > 0.05$; or intrusions, $F(2,21) = 1.284$, $p > 0.05$ (see Table 5). All ANCOVA results confirmed the ANOVA. The experimental instruction did not facilitate recall of either explicitly stated or inferable information.

-----------------------
Insert Table 5 about here.
-----------------------

Reading Test of the Stanford Achievement Test

One-way ANOVAs were performed on the raw scores from the reading comprehension and vocabulary subtests of the Stanford Achievement Test. In each case, treatment served as the single independent variable. Also, both ANCOVAs were employed.

For the reading comprehension scores, ANOVA results indicated a significant difference among the treatments, $F(2,21) = 7.773$, $p < 0.01$ (see Table 6).

-----------------------
Insert Table 6 about here.
-----------------------

The Newman-Keuls post hoc procedure computed at the .05 level of significance revealed that the performance of the Question group exceeded that of the other two groups, which did not differ from one another. However, the same post hoc procedure applied to the ANCOVA results indicated that the Strategy group also surpassed the Control group. Ironically, the weighted ANCOVA supported the ANOVA (although the Strategy/Control comparison was very close to reaching the value necessary to achieve significance at the .05 level). Thus, on this standardized comprehension test, the children receiving the Question intervention performed better than those in the other groups, and there is some reason to believe that the Strategy group exceeded the Control group.
ANOVA results indicated no differences on the vocabulary scores among the instructional groups, $F(2,21) = .665, p > .05$ (see Table 6). Both ANCOVAs supported this analysis. The vocabulary test is not a reading task, but an auditory assessment. Hence, the results tend to confirm an underlying equality among the three groups with respect to general world knowledge, thus strengthening the conclusion that the comprehension differences favoring the experimental groups were due to the intervention methods rather than to any world knowledge or general aptitude advantage.

Discussion

At the outset, we made five predictions about the results of this study that seemed reasonable to us in the light of recent theory and research regarding inferential processes. The data support all five of those predictions, albeit some to a greater degree than others.

First, we suggested that a technique which focused on helping children make connections between what they already know and what is in a text should increase the likelihood that they would draw inferences spontaneously. Clearly this occurred within the context of the instructional stories; the Strategy group was superior to the Control group on every comparison involving inferential measures. When scores were adjusted for significant pretest differences (the weighted ANOVA), the Strategy group also exceeded the practice only Question group. With respect to the experimenter-designed test, the Strategy group did not fare so well. In only one comparison out of four (post hoc tests based on the weighted ANCOVA for weighted posttest scores in the familiar condition) is there any reason to believe that Strategy treatment induced transfer, and even that is statistically marginal ($p < .066$).
The free recall data offer no support for any treatment. However, the results from the Stanford test, when the scores were adjusted for the pretest disadvantage of this group, tend to support its efficacy. It is commonplace to argue that achievement tests do not measure "higher level" comprehension tasks; however, the Stanford uses a preponderance of modified cloze items. We know of no way to determine which choice fits into a cloze blank save by reference to prior knowledge. Hence we are not surprised that this inferential treatment exhibited some transfer to the Stanford test. Some contradictory results notwithstanding, we believe the weight of evidence supports the efficacy of the Strategy treatment; we recognize, however, that the data better support an argument emphasizing its localized rather than its broadly transferable effects.

Our second prediction, based upon the engaged time on task argument, was that simply enabling children to practice answering inference questions would enhance their ability and inclination to do so spontaneously. Like that for the Strategy treatment, the data, while mixed, tend to support the efficacy of the Question treatment. On every comparison involving drawing inferences from the instructional stories, the Question treatment exceeded the Control. And on both adjusted and unadjusted comparisons for two of the three transfer tasks, the Question group exceeded the Control, and often the Strategy group. We acknowledge the fact that pretest measures tended to favor this group; however, their clear advantage even in the face of covariance analyses leads us to conclude that the treatment rather than their preexperimental advantage in ability accounted for the posttest differences favoring them.
The third prediction was that students trained in an inferencing set, whether by suggestion or sheer practice, would process the explicit message of the text better even than those students who were given a direct focus on that explicit message. This, we argued, would occur because inferential tasks, while they may have to be resolved by reference to prior knowledge, nonetheless require readers to use the text to acquire cues to direct them to particular schemata stored in memory. There are three sources of data to evaluate this prediction: the literal questions from the instructional stories, the literal questions from the transfer stories, and the recall of textual information in the free recall task. On the first of these measures, both experimental groups exceeded the Control; on the second and third, there were no significant differences among the groups. Whatever effects support this prediction, then, are highly localized within the context in which the treatments occurred. On the other hand, we would point out that an emphasis on inference never resulted in a decrement to the comprehension or recall of explicitly stated information.

Predictions four and five were disclaimers. Four suggested that whatever treatment effects emerged would be subject to a prior knowledge filter. Indeed, the posttest results on the transfer stories (the weighted ANCOVA for weighted-implicit scores) support the prediction quite directly. For the unfamiliar story there were no treatment differences; they emerged only on the familiar transfer story. Prediction five indicated that any growth in inferencing ability would not overcome the inherent advantage typically attributed to literal questions. The data from the transfer stories directly support that prediction; literal questions were easier. However, we would
point out that these comparisons are a weak test of the prediction because there is no control for comparability of content difficulty across item types. Even so, a recent study (Pearson, Hansen, & Gordon, in press) finds the same advantage even when content difficulty is controlled across item types.

In conclusion, three comments deserve emphasis. First, we are more impressed with the local than the transfer effects in this study. We wonder how distant from an instructional setting we can expect children to "spontaneously" apply learned strategies or behaviors. But replication of these treatments with older students—for whom greater maturity and experience might increase the likelihood of transfer—is necessary before the transferability issue can be adequately evaluated. Second, we underscore the ecological invalidity of our design. We did conduct the study in a school, using whatever instructional materials happened to be scheduled, and taking instructional groups the way they came to us; however, the teaching was done by one of the experimenters rather than the classroom teacher, and the non-treatment-related activities for each lesson were tightly controlled to mitigate against casual confounding between treatments. Tryouts in less controlled environments seem in order. Third, we believe that the preponderance of evidence in this study supports the efficacy of the experimental treatments. While we think that the treatments deserve larger scale tryouts in classroom situations by a variety of teachers working with a variety of children and materials, we are encouraged by results which suggest that both instruction and practice have direct consequences on children's comprehension performance.
Inference Training

Reference Notes


References


Pearson, P. D., & Camparell, K. B. *The comprehension of text structure.* In J. Guthrie (Ed.), *Reading comprehension and education.* Newark, Del.: International Reading Association, in press.


By emphasizing slot-filling inferences we do not mean to imply that this is the only kind of inference suggested by schema theory. Other kinds are not only possible but essential; they simply do not relate to the scope of this investigation. A second kind of inference involves what Trabasso calls text-connecting (see Trabasso, Nicholas, Omanson, & Johnson, Note 1, for examples). When a student recognizes that the action in sentence 13 caused the state described in sentence 15, he has connected two text segments via a causal link. Such inferences are common and essential in text comprehension. As we have discussed elsewhere (Pearson & Camparell, in press), such inferences require the invocation of the Gricean principle of cooperation between author and reader: No author places two sentences in close proximity to one another unless he or she is offering the reader an invitation, if not a license, to infer that the one explains, causes, enables, precedes, or embellishes the other. A third kind of inference is involved in the very process of schema instantiation. In example (1) in the article, the instantiation of a building schema involves an inference based upon the filling of a few variable slots specified in the story--carpenter, hit, nail.

In this instance we violated our general pattern of using both pretest scoring measures as covariates. We did so on the grounds that similarity of pre- and posttest measures should prevail as a criterion for adjusting post-test scores.
Table 1

Instructional Procedures Used with the Three Groups

<table>
<thead>
<tr>
<th>Day</th>
<th>Strategy Group</th>
<th>Question Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day One</td>
<td>Vocabulary Instruction</td>
<td>Vocabulary Instruction</td>
<td>Vocabulary Instruction</td>
</tr>
<tr>
<td>Day Two</td>
<td>Pre-reading Activity</td>
<td>Pre-reading Activity</td>
<td>Pre-reading Activity</td>
</tr>
<tr>
<td></td>
<td>Experimental Procedure</td>
<td>Basal-reader Procedure</td>
<td>Basal-reader Procedure</td>
</tr>
<tr>
<td>Day Three</td>
<td>Guided Reading</td>
<td>Guided Reading</td>
<td>Guided Reading</td>
</tr>
<tr>
<td></td>
<td>4:1 ratio of Literal: Inferential Questions</td>
<td>100% Inferential Questions</td>
<td>4:1 ratio of Literal: Inferential Questions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Experimental Procedure</td>
</tr>
<tr>
<td>Day Four</td>
<td>Phonic/Skill Activities</td>
<td>Phonic/Skill Activities</td>
<td>Phonic/Skill Activities</td>
</tr>
</tbody>
</table>
Table 2

Means and Standard Deviations for Comprehension Questions Following Instructional Stories

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Question Group</th>
<th>Strategy Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( M^a )</td>
<td>( SD^b )</td>
<td>( M^c )</td>
</tr>
<tr>
<td>Literal</td>
<td>5.75</td>
<td>1.16</td>
<td>5.62</td>
</tr>
<tr>
<td>Inferential</td>
<td>8.00</td>
<td>0.93</td>
<td>7.93</td>
</tr>
<tr>
<td>Strategy</td>
<td>7.13</td>
<td>2.17</td>
<td>6.93</td>
</tr>
<tr>
<td></td>
<td>7.00</td>
<td>2.00</td>
<td>7.10</td>
</tr>
<tr>
<td></td>
<td>8.13</td>
<td>1.25</td>
<td>8.18</td>
</tr>
<tr>
<td></td>
<td>8.63</td>
<td>1.06</td>
<td>8.77</td>
</tr>
<tr>
<td></td>
<td>3.50</td>
<td>0.93</td>
<td>3.53</td>
</tr>
<tr>
<td></td>
<td>6.50</td>
<td>1.07</td>
<td>6.52</td>
</tr>
<tr>
<td></td>
<td>6.00</td>
<td>1.93</td>
<td>6.04</td>
</tr>
</tbody>
</table>

\( ^a \) Reports unadjusted means.

\( ^b \) Standard deviations for unadjusted means.

\( ^c \) Means adjusted for total scores of experimenter-designed pretest.

\( ^d \) Means adjusted for weighted-implicit scores of experimenter-designed pretest.
### Table 3

<table>
<thead>
<tr>
<th>Group</th>
<th>Strategy</th>
<th>Control</th>
<th>Familiarity</th>
<th>Comprehension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unfamiliar: Literal</td>
<td>5.38</td>
<td>2.20</td>
<td>4.38</td>
<td>1.69</td>
</tr>
<tr>
<td>Unfamiliar: Inferential</td>
<td>4.38</td>
<td>1.41</td>
<td>3.63</td>
<td>1.30</td>
</tr>
<tr>
<td>Familiar: Literal</td>
<td>7.75</td>
<td>1.49</td>
<td>7.13</td>
<td>1.46</td>
</tr>
<tr>
<td>Familiar: Inferential</td>
<td>7.38</td>
<td>1.41</td>
<td>6.50</td>
<td>1.41</td>
</tr>
<tr>
<td>Total</td>
<td>24.88</td>
<td>21.63</td>
<td>36.40</td>
<td>21.13</td>
</tr>
</tbody>
</table>

Means and Standard Deviations for Experimenter-Designed Posttests
Table 4
Means and Standard Deviations for Weighted-Implicit Scores of Experimenter-Designed Posttest

<table>
<thead>
<tr>
<th></th>
<th>Question Group</th>
<th>Strategy Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M^a$ $SD^b$ $M^c$</td>
<td>$M^a$ $SD^b$ $M^c$</td>
<td>$M^a$ $SD^b$ $M^c$</td>
</tr>
<tr>
<td>Familiar</td>
<td>29.50 3.21 28.95</td>
<td>26.63 3.34 27.09</td>
<td>22.88 5.41 22.96</td>
</tr>
<tr>
<td>Unfamiliar</td>
<td>23.75 4.33 22.46</td>
<td>19.38 4.60 20.47</td>
<td>20.25 6.71 20.44</td>
</tr>
</tbody>
</table>

$^a$Reports unadjusted means
$^b$Standard deviations for unadjusted means
$^c$Means adjusted for weighted-implicit scores of experimenter-designed pretest
Table 5

Means and Standard Deviations for Categories of Free-Recall Statements

| Recall Category | Question Group | | | | Strategy Group | | | | Control Group | | |
|-----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                 | M<sup>a</sup> | SD<sup>b</sup> | M<sup>c</sup> | M<sup>d</sup> | M<sup>a</sup> | SD<sup>b</sup> | M<sup>c</sup> | M<sup>d</sup> | M<sup>a</sup> | SD<sup>b</sup> | M<sup>c</sup> | M<sup>d</sup> |
| Textual         | 3.50           | 2.00            | 3.59            | 3.50            | 3.00           | 1.20            | 2.93            | 3.00            | 3.00           | 2.00            | 2.98            |
| Scriptal        | 2.50           | 1.85            | 1.96            | 1.59            | 1.63           | 3.16            | 2.04            | 2.40            | 1.25           | 3.15            | 1.38            |
| Intrusions      | 1.50           | 2.14            | 1.38            | 1.36            | 1.38           | 1.51            | 1.47            | 1.50            | .38            | .52             | .40             |

<sup>a</sup> Reports unadjusted means
<sup>b</sup> Standard deviations for unadjusted means.
<sup>c</sup> Means adjusted for total scores of experimenter-designed pretest
<sup>d</sup> Means adjusted for weighted-implicit scores of experimenter-designed pretest.
<table>
<thead>
<tr>
<th>Subtest</th>
<th>Question Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehension</td>
<td>M₁</td>
<td>M₂</td>
</tr>
<tr>
<td></td>
<td>SD₁</td>
<td>SD₂</td>
</tr>
<tr>
<td></td>
<td>MC₁</td>
<td>MC₂</td>
</tr>
<tr>
<td></td>
<td>Md₁</td>
<td>Md₂</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>85.00</td>
<td>83.93</td>
</tr>
<tr>
<td></td>
<td>31.25</td>
<td>2.43</td>
</tr>
<tr>
<td></td>
<td>30.50</td>
<td>3.59</td>
</tr>
</tbody>
</table>

Notes:
- aReports unadjusted means
- bStandard deviations for unadjusted means
- cMeans adjusted for total scores of experimenter-designed pretest
- dMeans adjusted for weighted-implicit scores of experimenter-designed pretest

Inference Training
39
No. 1: Durkin, D. *Comprehension Instruction—Where are You?*, October 1977. (ERIC Document Reproduction Service No. ED 146 566, 14p., PC-$1.82, MF-$0.83)


No. 4: Jenkins, J. R., & Pany, D. *Teaching Reading Comprehension in the Middle Grades*, January 1978. (ERIC Document Reproduction Service No. ED 151 756, 36p., PC-$3.32, MF-$0.83)

No. 5: Bruce, B. *What Makes a Good Story?*, June 1978. (ERIC Document Reproduction Service No. ED 158 222, 16p., PC-$1.82, MF-$0.83)

No. 6: Anderson, T. H. *Another Look at the Self-Questioning Study Technique*, September 1978. (ERIC Document Reproduction Service No. ED 163 441, 19p., PC-$1.82, MF-$0.83)


No. 8: Collins, A., & Haviland, S. E. *Children's Reading Problems*, June 1979. (ERIC Document Reproduction Service No. ED 172 188, 19p., PC-$1.82, MF-$0.83)

No. 9: Schallert, D. L., & Kleiman, G. M. *Some Reasons Why Teachers are Easier to Understand than Textbooks*, June 1979. (ERIC Document Reproduction Service No. ED 172 189, 17p., PC-$1.82, MF-$0.83)

No. 10: Baker, L. *Do I Understand or Do I not Understand: That is the Question*, July 1979. (ERIC Document Reproduction Service No. ED 174 948, 27p., PC-$3.32, MF-$0.83)


No. 13: Adams, M., & Bruce, B. *Background Knowledge and Reading Comprehension*, January 1980.


CENTER FOR THE STUDY OF READING

TECHNICAL REPORTS

No. 1: Halff, H. M. Graphical Evaluation of Hierarchical Clustering Schemes, October 1975. (ERIC Document Reproduction Service No. ED 134 926, 11p., PC-$1.82, MF-$0.83)


No. 3: Goetz, E. T. Sentences in Lists and in Connected Discourse, November 1975. (ERIC Document Reproduction Service No. ED 134 927, 75p., PC-$4.82, MF-$0.83)

No. 4: Alessi, S. M., Anderson, T. H., & Biddle, W. B. Hardware and Software Considerations in Computer Based Course Management, November 1975. (ERIC Document Reproduction Service No. ED 134 928, 21p., PC-$1.82, MF-$0.83)


No. 7: Ortony, A. Names, Descriptions, and Pragmatics, February 1976. (ERIC Document Reproduction Service No. ED 134 931, 11p., PC-$1.82, MF-$0.83)

No. 8: Mason, J. M. Questioning the Notion of Independent Processing Stages in Reading, February 1976. (Journal of Educational Psychology, 1977, 69, 288-297)


No. 15: Schwartz, R. M. Strategic Processes in Beginning Reading, November 1976. (ERIC Document Reproduction Service No. ED 134 937, 19p., PC-$1.82, MF-$0.83)

No. 16: Jenkins, J. R., & Pañy, D. Curriculum Biases in Reading Achievement Tests, November 1976. (ERIC Document Reproduction Service No. ED 134 938, 24p., PC-$1.82, MF-$0.83)


No. 20: Kleiman, G. M. The Effect of Previous Context on Reading Individual Words, February 1977. (ERIC Document Reproduction Service No. ED 134 941, 76p., PC-$6.32, MF-$0.83)


No. 98: Green, G. M. *Discourse Functions of Inversion Construction*, July 1978. (ERIC Document Reproduction Service No. ED 160 998, 42p., PC-$3.32, MF-$0.83)


No. 105: Ortony, A. * BeyondLiteral Similarity*, October 1978. (ERIC Document Reproduction Service No. ED 166 635, 58p., PC-$4.82, MF-$0.83)


No. 115: Gearhart, M., & Hall, W. S. *Internal State Words: Cultural and Situational Variation in Vocabulary Usage*, February 1979. (ERIC Document Reproduction Service No. ED 165 131, 66p., PC-$4.82, MF-$0.83)


No. 120: Canney, G., & Winograd, P. *Schemata for Reading and Reading Comprehension Performance*, April 1979. (ERIC Document Reproduction Service No. ED 169 520, 99p., PC-$6.32, MF-$0.83)


No. 124: Spiro, R. J. *Etiology of Reading Comprehension Style*, May 1979. (ERIC Document Reproduction Service No. ED 170 734, 21p., PC-$1.82, MF-$0.83)


No. 147: Stein, N. L., & Goldman, S. Children's Knowledge about Social Situations: From Causes to Consequences, October 1979.


