DOCUMENT RESUME

ED 237 578 TM 830 848

AUTHOR Baker, Linda
TITLE Spontaneous versus Instructed Use of Multiple Standards for Evaluating Comprehension: Effects of Age, Reading Proficiency, and Type of Standard.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE [83]
GRANT NIE-G-81-0100
NOTE 41p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Age Differences; Cognitive Style; *Evaluation Criteria; Grade 4; Grade 5; Intermediate Grades; *Prose; Reading Ability; *Reading Comprehension; *Reading Strategies; *Self Evaluation (Individuals)
IDENTIFIERS *Comprehension Monitoring; Embedding (Grammar); *Embedding Transformations

ABSTRACT

Fourth and sixth grade children differing in reading proficiency read and commented on brief expository passages containing three different types of embedded problems (nonsense words, prior knowledge violations, and internal inconsistencies). Half of the children were specificially instructed as to the types of standards they should apply in order to detect the problems (lexical, external consistency, and internal consistency); the remaining children were simply instructed to look for problems. Both quantitative and qualitative differences in standard use were revealed by the children's comments about all parts of the passages. Older and better readers used more different standards and they used them more frequently than younger and poorer readers. The lexical standard was more likely to be adopted spontaneously than the other two standards and it was the only standard used by a substantial proportion of both younger and poorer readers. The results demonstrate that children differ in their ability to decide for themselves whether or not they understand but that their performance depends in part on the amount of guidance they are given. (Author)

Spontaneous Versus Instructed Use of Multiple Standards for Evaluating Comprehension: Effects of Age, Reading Proficiency, and Type of Standard

Linda Baker

University of Maryland Baltimore County

Running head: Standards of Evaluation

Abstract

Fourth and sixth grade children differing in reading proficiency read and commented on brief expository passages containing three different types of embedded problems (nonsense words, prior knowledge violations, and internal inconsistencies). Half of the children were specifically instructed as to the types of standards they should apply in order to detect the problems (lexical, external consistency, and internal consistency); the remaining children were simply instructed to look for problems. Both quantitative and qualitative differences in standard use were revealed by the children's comments about all parts of the passages. Older and better readers used more different standards and they used them more frequently than younger and poorer readers. The lexical standard was more likely to be adopted spontaneously than the other two standards and it was the only standard used by a substantial proportion of both younger and poorer readers. The results demonstrate that children differ in their ability to decide for themselves whether or not they understand but that their performance depends in part on the amount of guidance they are given.

A common paradigm in research on comprehension monitoring has been to examine children's ability to identify a particular type of problem embedded in a text (e.g., Garner, 1980, Garner & Kraus, 1981-82; Harris, Kruithof, Terwogt, & Visser, 1981; Markman 1979). A difficulty with this approach is that it has fostered a conception of comprehension monitoring as a unitary phenomenon rather than as a multidimensional process. Consider, for example, a typical study in which children are tested for their detection of contradictory information embedded within a passage. If the children fail to notice the inconsistencies, the researcher concludes that the children were poor at monitoring their comprehension. What the study in fact has shown, however, is that the children were poor at evaluating their understanding with respect to internal consistency. In other words, the study reveals a difficulty in using one particular standard of evaluation; it says nothing about the use of other standards.

The purpose of this paper is to present evidence for a richer conceptualization of comprehension monitoring. Just as we have come to realize that comprehension is a complex multi-faceted process (cf. Spiro, Bruce, & Brewer, 1980), so too must we realize that effective evaluation of that comprehension is multidimensional (cf. Baker, in press; Markman, 1981). There are many different standards or criteria that readers must take into account when they decide whether or not they understand, a fact that much of the existing research has failed to recognize. Moreover, there may be differences in the likelihood that particular standards will be applied, differences which may account for some of the apparent contradictions in the literature. For example, one study may show that ten year olds are

exceptionally good at noticing embedded nonsense words while another may show that ten year olds are very poor at noticing inconsistencies. In both studies, general conclusions are drawn about comprehension monitoring as a unitary phenomenon; in one case comprehension monitoring is concluded to be good, in the other poor. The discrepancy can be better understood when we consider that two different standards of evaluation were required.

The study to be reported examined children's use of multiple standards of evaluation while reading expository text. Three standards were selected for investigation: lexical, internal consistency, and external consistency. The lexical standard involves consideration of individual word meanings and can be applied without regard to surrounding context. The internal consistency standard involves evaluation of the consistency of different propositions within the text. The external consistency standard involves evaluation of the consistency of a proposition in the text with respect to prior knowledge. The standards obviously differ in their processing demands and so are likely to differ in their ease of application. For example, use of the internal consistency standard entails several steps: accessing a memory representation if the propositions are widely separated, integrating the propositions, and comparing them. Use of the lexical standard, in contrast, requires only a check that the word is present in one's internal lexicon.

Problems were embedded within the passages to ensure that there would be at least some occasions when each type of standard could be used. Accordingly, some passages contained nonsense words, some contained internal inconsistencies and some contained prior knowledge violations. But there was no assumption that these problems were the only aspects of the passages that

[illegible] be [illegible] problematic by individual children. Although problem [illegible]ntification served as one dependent measure; of more interest were the types of standards subjects used and their patterns of use. Therefore, all of the comments made about the entire passages were coded as to the type of standard they reflected and these data provided the basis for several additional dependent measures.

Participants in the study were fourth and sixth graders, identified as either better or poorer readers on the basis of standardized test scores. The children's task was to read each passage silently, underlining anything that appeared problematic and then explaining why they had done so. Half of the children were specifically told that three different types of problems would be present in the passages and they were given examples of each type. The remaining children were simply informed that problems would be present. The specific instruction condition was designed to provide information about the use of the different standards when subjects are induced to use them. The general instruction condition was designed to reveal what types of standards the subjects spontaneously adopt when given instructions to evaluate their understanding carefully. Note that this condition does not correspond to an uninformed control group: There is too much evidence that readers frequently will not identify embedded problems if they are not told that problems are present, in line with Grices's (1975) cooperative principle.

It was expected that the specificity of the instructions would have a strong effect on overall standard use, consistent with previous evidence that children identify more problems when explicitly told what to look for (e.g., Baker, Note 1; Markman & Gorin, 1981). Of particular interest was whether the

instruction manipulation would have differential effects depending on age, reading level, and type of standard. For example, perhaps older and/or better readers spontaneously adopt several different standards for evaluating their comprehension. If so, there should be little difference between general and specific instructions for the more mature readers. Alternatively, perhaps the poorer readers will have difficulty applying the standards when they are told to do so, so only the better readers will benefit from specific instructions. It is also possible that certain standards are more likely to be adopted spontaneously than others. A likely candidate is the lexical standard which, because it is one that children are frequently exhorted to apply, may be used equally often by children receiving general and specific instructions.

Subjects

A total of 108 children participated in the study. The children were enrolled in the fourth and sixth grades of three suburban public schools. There were 54 fourth graders (29 girls), with a mean age of 9.46 years (s.d.=.37). There were 54 sixth graders (27 girls), with a mean age of 11.51 years (s.d.=.40). Candidates for participation in the study were pre-selected on the basis of their reading scores on the California Achievement Test. Children were selected for the better reading group if they had stanine scores of 7, 8, or 9 and for the poorer reading group if they had stanine scores of 3, 4, or 5. (Although a score of 5 is normatively average based on the entire U.S. sample, within this particular county, the mean was 6.5.) The final sample consisted of 31 better and 23 poorer fourth grade readers (mean stanine scores = 7.77 and 4.09, s.d. = .76 and .75, respectively), and 32 better and 22 poorer sixth grade readers (mean stanine scores = 7.38 and 4.26, s.d. = .49

and .70, respectively). The two sexes were evenly represented within each group. The subjects within groups were randomly assigned to one of two instructional conditions, with the constraint that there be equal numbers of boys and girls in each condition. (Note that gender was not treated as a factor in this study because the number of subjects per cell would have been too small.)

Materials

The first step in developing the passages was to consult three non-fiction books written for elementary school children. One book dealt with the weather, a second with the planets, and the third with geographical regions in the United States. The books served as the source of information for the passages, but the passages bore little structural similarity to the books. Five passages were constructed dealing with each topic. The passages were very similar in overall length (mean number of words = 49.25, s.d. = 1.42, range = 47-51) and all consisted of six sentences. Across passages, the length of each sentence occupying the same serial position was the same (i.e., sentence number one had eight words, sentence 2 had 10 words, etc). The opening sentence of each passage introduced the topic and subsequent sentences provided relevant facts about the topic. The number of propositions was comparable across passages (mean = 11.96, s.d. = 1.16). The passages were checked for readability using the Harris-Jacobsen formula (Harris & Sipay, 1980). This formula, which takes both sentence length and vocabulary difficulty into account, yielded a mean readability level of 3.22 (s.d. = .13, range = 3.03 to 3.43).

After the passages were written, 12 of the passages were made problematic by replacing a single word with another. Four of the modified passages contained two-syllable nonsense words which followed standard rules of English orthography. An example of a nonsense word appears in the following sentence: "It is so ho that most brugens would melt there." Four other passages presented information that violated common world knowledge. An example is: "They used sand from the trees to make many things." The third type of problem, embedded in the remaining four passages, was internal inconsistency, created by making one sentence in the passage conflict with a previously presented sentence. One sentence intervened between these two target sentences. An example follows: "The temperature on Venus is much higher than boiling water. Venus is about the same size as Earth. But it is much too cold for us to live there." The nonsense words and the prior knowledge violations were always embedded within the middle of the fourth sentence and they always replaced nouns. The internal inconsistencies always involved information contained in sentences two and four; the substitution appeared in the middle of sentence 4 and was always an adjective.

In order to ensure that the problems would be perceived as such by adult readers, the passges were first presented to 20 undergraduates who were asked to identify the problems. They were explicitly informed that three different types of problems would be present in the passages and were given examples of each type. Subjects correctly identified 96% of the nonsense words, 87% of the inconsistencies, and 85% of the prior knowledge violations. In a few instances, several students were in agreement that there was a problem in a sentence that was not intended to be problematic and so these sentences were rewritten.

The materials were then subjected to a second screening test with a different group of 20 undergraduates who were presented with versions of the passages that had the critical word deleted (i.e., the word that changed the passage from non-problematic to problematic). The subjects were instructed to try to fill in the missing word by using the surrounding context. Of concern was whether the context would constrain the word choices to the same extent for the different problem types, since differences in detection could result if the appropriate words were highly constrained for one type (and so subjects need not process the word very deeply), but not for others. However, contextual constraints were high for all problem types, with subjects supplying the correct word or a reasonable substitution 95-100% of the time.

The final versions of the passages were assembled into booklets. Each passage was typed on a separate sheet of paper and was headed by a descriptive title. At the bottom of each page was a set of four schematic faces with expressions ranging from very happy (full smile) to very sad (full frown). These faces were used by the children to indicate whether the passage was problematic, and if so, to what extent it affected comprehensibility (see Procedure section). The same three passages appeared in first, second, and third positions in each booklet and contained a nonsense word, an inconsistency, and a prior knowledge violation, respectively. These passages served as warm-up passages (i.e., although no feedback was provided, children's responses were not scored). The remaining 12 passages were arranged in a constrained, partially randomized, order. The constraint was that a particular problem type had to appear in a particular position across booklets, but the passage containing that problem was randomly selected from

the pool. For example, the first experimental passage was nonproblematic; for any given child it could be any one of the three nonproblematic passages. The second passage always contained a prior knowledge violation, and it too could be any one of three, etc.

Procedure

All subjects were seen individually. In order to preclude the possiblity of inadvertent bias, the experimenter was "blind" as to the subject's reading level during testing. At the beginning of each session, the children were given a booklet containing the passages and were asked to fill in their names, ages, and dates of birth on the cover sheet. At the bottom of the page were four schematic faces that the children were asked to refer to while they listened to the relevant portions of the instructions.

The instructions were presented on tape, recorded by the experimenter. Children in both conditions were instructed that their task was to try to find the problems in some short passages dealing with things they learn about in school. A problem was defined as "something that might confuse people or something that people might have trouble understanding." Children in the specific instruction condition were given further information about the exact nature of the problems and two examples of ea type. The examples of the internal inconsistencies and prior knowlege violations were the same as those used by Markman and Gorin (1981); the nonsense word examples followed the same format. The terms used to describe the three problem types were: 1) two parts of the paragraph that don't make sense together; 2) things that aren't true; and 3) words that aren't really words.

Children in both conditions were instructed to underline anything they thought was a problem as they were reading. Then, when they finished, they were to rate the comprehensibility of the passage by circling one of the faces at the bottom of the page. If they circled the face with the big smile, that meant the passage did not have any problems and was easy to understand. The face with the small smile meant there might have been a little problem; but the passage was still easy to understand. The face with the small frown meant there was a little problem and the passage was a little hard to understand. And the face with the big frown meant there was a big problem, which made the passage very hard to understand. (As it turned out, subjects rarely circled either of the frowning faces; instead, they used the big smile to indicate no problem and the small smile to indicate a problem. For this reason, the rating data provided by the faces were not subjected to any statistical analyses.) The subjects were further informed that after they circled a face, they were to explain why they made that choice and also why they underlined any words or phrases.

After the instructions were presented, the children were asked to turn to the first (practice) passage and read it silently to themselves. Care was taken to ensure that the children understood the task by reminding them, if necessary, to underline problematic information as soon as they encountered it and by pointing out any lack of correspondence between ratings and underlining (e.g., the face with the small smile was circled but no sentences were underlined). The children were not given feedback as to whether or not they had correctly identified the intended problems. After completing the three practice passages, the children went on to the 12 experimental passages. For

each passage, the children read it silently at their own pace, underlining any problematic sections as they went along. When they finished reading the passage, they circled one of the faces and then explained why they had done so.

Upon completion of the task, the children were asked a standard set of questions designed to reveal how they had interpreted any problems they did not report. Before asking each question, the experimenter turned to the appropriate passage and placed it in front of the child, who was encouraged to reread it to find the answer. If the child did not spontaneously indicate awareness of the problem after answering the question, the experimenter asked, "Does that make sense to you?" If the child still did not indicate problem awareness, a second prepared question was asked. To illustrate, the questions for the prior knowledge problem embedded in the sentence, "They used sand from the trees to make many things" were as follows: What part of the tree did the settlers use for making things? and Does sand come from trees? The questions were only asked for those problems the child did not initially identify, so the number of questions asked varied from child to child.

All sessions were tape recorded and the tapes were subsequently transcribed. The length of the sessions varied, averaging about 25 minutes per child, with a range of 10 to 45 minutes.

Scoring

Problem Identification. Responses were scored as problem identifications if the child underlined the target information and gave an adequate explanation of the nature of the problem. It was not necessary for the child to specify the problem type. So, for example, if a child underlined the

nonsense word brugens and then said, "I don't know what that word means," the response was scored as a correct identification. Similarly, if a child explained that she underlined the phrase, "sand from trees" because she didn't think you could get sand from trees, this was scored as correct. There was never any question on the basis of the responses to nonsense words and prior knowledge violations as to whether the child noticed the intended problem. However, there was occasionally some ambiguity in the initial response to internal inconsistencies which necessitated further questioning. Most of the children only underlined the second target sentence of the inconsistency, and when asked to explain why they underlined that sentence, some did not mention the first target sentence. For example, if a child underlined Venus is much too cold for us to live there and explained only that he knew Venus wasn't too cold, it was necessary to ask him how he knew that. At this point, most of the children identified the relevant information in the first target sentence. However, on 24 of the 324 possible occasions (3 inconsistencies x 108 subjects), the children indicated that they knew something was not true because, for example, they had learned it in school. These children were then asked if there was anything in the passage that supported this idea. If they mentioned the first target sentence, they were given credit for identifying the inconsistency, but if they did not, their responses were not scored as correct. Five percent of the responses fell in this latter category.

Standard application. Once the response protocols were fully transcribed, the children's responses to "non-problematic" segments of text were coded as to the type of standard they revealed. The coding scheme used was that established by Baker (in press) and included seven categories: the three

which were the main focus of the present study: lexical, external consistency, and internal consistency: plus syntax; informational clarity and completeness (e.g., "They should tell you more about how you get medicine from tree bark"); propositional cohesiveness (e.g., "Does the word "it" refer to the desert or the weather?") and structural cohesiveness (e.g., Here it says it's hot and then it says it again down here; why are they repeating themselves?"). It is important to note that no judgments were made as to whether the comments reflect valid problems. For example, if a subject misunderstood a particular sentence and so said that the following sentence was inconsistent, this was scored as an application of the internal consistency standard just as if the two sentences were in fact inconsistent.

The protocols were scored by two independent judges who were able to classify all but 6% of the comments into one of the seven categories. Inter-judge agreement was high and the few discrepancies were resolved through discussion. However, the latter four standards were seldom used; only the informational completeness standard was reported by more than four children (14 children used this standard). Most of the data analyses therefore will focus only on the three most commonly used standards: lexical, internal consistency and external consistency.

Responses to follow-up questions. Children's responses were scored for the extent to which they revealed problem awareness and/or resolution. A child received a score of 1 if he or she spontaneously mentioned, on first being questioned, that there was a problem. A score of 2 was awarded if the child reported the problem when asked, "Does that make sense to you?", and a score of 3 was given if the child acknowledged the problem when being asked

directly about the problematic information. The maximum score of 4 was given on the rare occasions when the child still failed to see the problem.

Results

This section of the paper is divided into three sections. The first section presents the results for the problem identification task. The data analyses included both analysis of variance and multiple regression procedures. The second, most important section focuses on the application of the various standards throughout the entire testing session. Several different dependent measures will be examined: 1) frequency of use; 2) failures to use particular standards; 3) exclusive use of single standards; 4) number of different standards used; and 5) the relationship of standard use to problem identification. The third section presents the results of the analysis of the subjects' responses to the follow-up questions.

Problem Identification

The mean number of problems of each type identified by the children as a function of grade, reading proficiency, and instruction condition is presented in Table 1. Note first of all the low levels of identification of all three

Insert Table 1 about here

problem types. These levels were affected, however, by each of the factors of interest, as revealed by a mixed-model analysis of variance. Sixth graders identified more problems than fourth graders (53% vs. 34%), $F(1,92)=14.44$, $p<.001$ and better readers identified more problems than poorer readers (54% vs. 29%), $F(1,92)=34.19$, $p<.001$. Children receiving specific information about

the nature of the problems identified more problems than children told only generally that problems would be present (55% vs. 32%), $F(1,92)=26.81$, $p<.001$. Finally, children identified more nonsense words than either prior knowledge violations or internal inconsistencies (53% vs. 39% and 38%, respectively), $F(2,184)=6.95$, $p=.001$.[1]

Contrary to expectations, problem type did not interact with grade, reading proficiency, or instruction condition. However, there were reliable interactions of reading proficiency with age, $F(1,92)=7.31$, $p<.01$, and of reading proficiency with instruction condition, $F(1,92)=4.36$, $p<05$. Although sixth graders identified more problems than fourth graders, this effect was largely attributable to children in the better reader group. The older better readers identified 68% of the problems as compared to the 40% identified by the younger better readers. Among the poorer readers, the sixth graders identified 31% of the problems and the fourth graders 26%, a nonsignificant difference. Thus, although there is substantial improvement with development among children who are effective readers, the less effective readers do not show significant gains. This pattern is consistent with the conventional wisdom that poorer readers tend to fall further behind as they go through school.

Children receiving specific instructions identified mnore problems than those receiving general instructions, but this effect too was mediated by reading level. Among the better readers, problem identification went from 39% with general instructions to 70% with specific instructions. The improvement for the poorer readers was much less substantial (23% to 34%), though still statistically reliable. Thus, the better readers were much more successful at

adopting the experimenter-provided criteria for evaluating the passages. This suggests that the difficulty poorer readers experience in evaluating their understanding is not simply the result of their not knowing what criteria to use. Nevertheless, the fact that they did show modest gains is encouraging; in fact, their identification rate under specific instructions was not significantly different from that observed for the better readers under general instructions.

The analysis of variance was based on a dichotomous classification of the children as being better or poorer readers. The reliable main effect of reading proficiency indicates that the two groups did indeed differ. But it does not indicate how much of the variance in problem identification is attributable to reading proficiency. To answer this question, multiple regression analyses were carried out using the subjects' actual stanine scores as predictor variables. A second predictor variable was the subject's age in years and months. Because instruction condition was a qualitative variable, the data were analyzed separately for the two conditions. The total number of problems identified served as the dependent variable.

We will consider first the regression analysis for subjects in the specific instruction condition. The predictor variables were entered into the regression equation through a forward stepping algorithm. Stanine score was the first variable to enter the regression equation, (F to enter = 25.05) accounting for 33% of the variance (r=.57). The age variable entered the equation on the second step (F to enter = 3.91). The multiple correlation was .62 and the combined proportion of the variance accounted for by the two variables was .38. The analysis indicates, then, that when subjects were

specifically instructed as to the types of problems they should seek, reading proficiency was a much stronger prediction of problem detection than chronological age.

The regression equation for subjects receiving general instructions yielded a rather different solution. The first variable to enter the equation was age (F to enter = 11.33), accounting for 18 percent of the variance (r=.42). Stanine score entered second (F to enter = 7.07), accounting for an additional 10% of the variance. The multiple correlation was .52 and the multiple r-square was .28. This analysis indicates that age is a better predictor of problem identification than reading proficiency when subjects are left to select the standards on their own. However, the two variables together account for less of the variance in performance than they do for subjects who received specific instructions.

Application of Standards Throughout the Testing Session

As noted earlier, the introduction of problems into passages is one way to assess readers' use of different standards of evaluation. However, the underlying assumption is that effective readers routinely apply certain standards to evaluate their understanding; therefore, the use of these standards can be revealed through any evaluative comment made about the text. This section of the paper will present information gleaned from analysis of the complete response protocols.

Frequency of application. Table 2 presents the mean number of times children applied the lexical, external consistency, and internal consistency standards. These data reflect standard application both in the service of identifying the various problems and in comments about text that was not

intended to be problematic. Remember, no judgment was made as to whether the standard was used appropriately or inappropriately from an adult perspective.

Insert Table 2 about here

Either situation would reveal that the child is evaluating her understanding with respect to a particular standard. An analysis of variance was carried out, with age, reading proficiency, and instruction conditon as between subjects factors and type of standard as a within-subjects factor.

The analysis revealed that neither the main effects of age nor reading proficiency were reliable (F's < 1.0) but the two factors interacted, $F(1,92)=5.29$, $p<.05$. This interaction differs from the age by proficiency interaction reported earlier for problem detection in the cell corresponding to fourth grade poorer readers. Children in this group applied the standards more frequently than either the fourth grade better readers or the sixth grade poorer readers (means = 2.84 vs. 2.18 and 2.24, respectively). In fact, their frequency of standard application did not differ from that of the sixth grade better readers (mean = 3.09). There is a difference, of course, in the effectiveness with which the standards were applied, as the fourth grade poor readers identified fewest actual problems. The present data indicate, however, that their low levels of problem identification cannot be attributed to such factors as an unwillingness to criticize the material or admit ignorance.

Children receiving specific instructions applied the three standards more frequently than those receiving general instructions, $F(1,92)=22.96$, $p<.001$.

A reliable grade by instruction condition interaction, $F(1,92)=3.92$, $p=.05$, indicated that fourth and sixth graders did not differ in the mean number of standard applications under specific instructions (3.46 vs. 3.27), but the older children spontaneously applied the standards more often under general instructions than did the younger (2.26 vs. 1.47). Recall that grade did not interact with instruction condition in the analysis of problem identification; both fourth and sixth graders showed comparable improvements from general to specific instructions, with the sixth graders better overall. Thus, the present data indicate that the fourth graders complied with task demands to use the standards of evaluation, but they were applied less effectively than were those used by the sixth graders.

The three standards were applied with different frequency, $F(2,184)=17.88$, $p<.001$. The most frequently used standard was that of external consistency (mean =3.63), next was the lexical standard (mean = 2.74), and least frequently used was the internal consistency standard (mean = 1.44). All differences between means were statistically reliable. These figures indicate that the external consistency standard in particular was applied in many situations other than those intended by the experimenter. (Recall that prior knowledge problems were detected at the same rate as inconsistencies, and both were less often identified than nonsense words.) An interaction of standard type with reading proficiency shows an interesting crossover effect, $F(2,184)=3.33$, $p<.05$. Whereas the poorer readers used the lexical and external consistency standards more often than the better readers, they used the internal consistency standard dramatically less often. Thus, although the problem detection data did not yield a reliable problem type by proficiency

interaction, the present data do indicate that poorer readers are much less likely to evaluate text for internal consistency than they are to evaluate for either external consistency or word understanding.

The frequency with which the different types of standards were used also varied with the nature of the instructions, $F(2,184)=4.84$, $p<.01$. Children specifically instructed to apply the standards used the external consistency and internal consistency standards more than twice as often as children who received only general instructions. The data also indicate that children are more likely to adopt external consistency and lexical standards when required to select their own criteria for evaluating text than they are to adopt internal consistency standards.

Failures to use particular standards. Additional information about differences in children's standard use was obtained by classifying the subjects as to whether they ever used a specific standard or not. The data base for this classification was again the total number of standards used, not simply those involved in identifying a problem. The classification is lenient in that it is based on the assumption that a single instance of standard use indicates that the standard is available in the child's repertoire. Table 3 shows the proportion of subjects who never used a specified standard. Visual inspection of the table makes it quite clear that these proportiond differed considerably across cells. In order to examine these differences more

Insert Table 3 About Here

systematically, separate multiway frequency tables were created for each of the three types of standards and tests of association were carried out using a log-linear model.

Let us consider the lexical standard first. Overall, the proportion of children who never used the standard is small, (.18) as one would expect given the relatively good identification of nonsense words. Tests of association revealed that more subjects in the general instruction condition never used the standard than subjects in the specific ($\underline{X}^2$=7.75, $\underline{p}$<.01) However, an interaction with reading proficiency showed that poorer readers were less likely to use the standard under specific irstructions, ($\underline{X}^2$=5.01, $\underline{p}$<.05). Additionally, a grade by reading proficiency interaction indicated that a substantial proportion of sixth grade poorer readers never used the standard, ($\underline{X}^2$=3.75, $\underline{p}$=.05). This latter finding is intriguing because it suggests that poorer readers become less willing to acknowledge word level comprehension problems as they grow older, having learned, perhaps, that there is a stigma associated with such admissions. Even when specifically told that nonsense words would be present, close to a third of the children failed to identify a single word as problematic.

Consider now the cell frequencies for children who never used an external consistency standard. Tests of association revealed a very strong effect of instruction ($\underline{X}^2$=23.71, $\underline{p}$<.001); not surprisingly, more subjects failed to use the standard if they were not specifically told to evaluate for external consistency. In addition, there were many more fourth graders who never adopted the standard than sixth graders ($\underline{X}^2$=14.35, $\underline{p}$<.001). None of the other tests of association were reliable. In particular, better readers were no more or less likely to use the standard than poorer readers.

With respect to the internal consistency standard, we find that grade, reading proficiency, and instructions all influence the likelihood of

adoption. More children used the standard under specific instructions than general, as one would expect ($\underline{X}^2$=14.10, $\underline{p}$<.001). More better readers used the standard than poorer readers ($\underline{X}^2$=14.38, $\underline{p}$<.001); and more older children used it than younger ($\underline{X}^2$=7.11, p<.001). None of the higher order interactions were reliable.

Finally, some comments about Table 3 as a whole can be made on the basis of visual inspection. First, note that among the sixth grade better readers who received specific instructions, there was not a single instance of failure to apply any standard at least once. No other groups showed this pattern. Among the fourth grade better readers with specific instruction, only the lexical standard was applied by all students. Note also the differences across standards. Overall, 18% of the children never used a lexical standard, 33% never used an external consistency standard, and a full 45% never used the internal consistency standard. The relative ordering suggests that lexical standards are more likely to be adopted than external consistency standards, which in turn are more likely to be adopted than internal consistency standards.

Exclusive use of a single standard. The proportion of children who used only one particular standard of evaluation throughout the testing session is shown in Table 4. With the exception of one child, the internal consistency standard was never used exclusively. This is consistent with the view that it is a relatively more sophisticated standard and hence is unlikely to be the only one available in a child's repertoire. The external consistency standard similarly was rarely used exclusively. Fewer than 5% of the children did so, all of whom, interestingly, were less effective readers. The lexical

standard, in contrast, was used exclusively by a substantial proportion (.28) of the subjects. The incidence differed considerably across cells and so tests

Insert Table 4 About Here

of association using a log-linear model were carried out. Results revealed that exclusive use of the lexical standard was more frequent in the general instruction condition than the specific (X^2=10.25, $p<.001$), that it was more frequent among poorer readers than better (X^2=6.45, $p<.01$; and that it was more frequent among the fourth graders than the sixth graders (X^2=6.04, $p<.01$). None of the higher order interactions showed significant associations. Thus, these findings suggest that younger and poorer readers are more likely to evaluate their understanding at the word level only, an outcome consistent with other studies suggesting over-reliance on a lexical standard (e.g., Garner, 1981).

Number of different standards applied. An indication of the variety of standards in a child's repertoire is provided by analysis of the number of different standards used. For this analysis we again consider all of the responses the children made, classified as to the type of standard they revealed. The coding scheme identifies a maximum of seven different standards; in actuality, no child used more than five. Table 5 shows the mean number of different standards used by the children as a function of grade, reading proficiency and instruction condition. An analysis of variance with these three between-subjects factors revealed reliable main effects of each, as well as a grade by instruction interaction. Sixth graders used more different standards than the fourth graders (2.5 vs 2.0), $F(1,92)=8.58$, $p<.01$. Better readers used more different standards than poorer readers (2.57 vs

1.93), $F(1,92)=14.44$, $p<.001$. And children receiving specific instructions not surprisingly used more different standards than those not given specific instructions (2.72 vs 1.79), $F(1,92)=30.19$, $p<.001$. Finally, the grade by instruction condition interaction, $F(1.92)=5.68$, $p<.05$, reflects the fact that fourth graders who received specific instructions did not differ reliably from the sixth graders who received specific instruction; however, fourth graders who received general instructions spontaneously adopted fewer different standards than did their sixth grade counterparts.

Insert Table 5 About Here

Relationship of standard use to problem detection. It has been assumed that any comments reflecting the use of a particular standard indicate that that standard is in the child's repertoire and that she can use it effectively. If this is true, then use of a particular standard should be accompanied by detection of at least some of the corresponding problems. If a child used a standard but did not detect any of the problems, this could indicate that the child was simply responding to demand characteristics of the task. The proportion of children who used a particular standard at least once but did not identify any of the corresponding problems was calculated. In almost half of the 24 cells, the proportion was 0, and in all but two, the proportion was .1 or less, indicating that the standards typically were used productively. The remaining two cells correspond to the use of the external consistency standard by poorer readers receiving specific instructions. Overall, 42% of the fourth graders and 20% of the sixth graders in these groups challenged the truth of various passage statements but did not identify any of the prior knowledge violations. This pattern, which serves to explain

the discrepancy mentioned earlier between problem detection and standard use, indicates that these less successful readers attempted to comply with the task demands but could not apply the external consistency standard effectively enough to identify the intended problems.

Responses to Follow-Up Questions

Children's responses to the questions they were asked about missed problems were scored as described in the Method section. Each child's mean score, averaged over problem types, was entered into an analysis of variance with grade, reading proficiency, and instruction condition as between-subjects factors. The only reliable effects were for grade, $F(1,83)=16.69$, $p<.001$ and reading proficiency $F(1,83)=11.98$, $p<.001$.[2] It made no difference whether the initial instructions had been general or specific. Sixth graders had lower scores than fourth graders, (1.77 vs 2.10), indicating that they perceived the nature of the problems more quickly. Similarly, better readers had lower scores than poorer readers (1.79 vs 2.07).

The fact that the younger and poorer readers had trouble perceiving the nature of the problems even when they were directly confronted with the relevant information is consistent with results reported by Garner and Taylor (1982). It suggests that problem identification is influenced by factors other than reading experience and proficiency. For example, logical reasoning skills seem to play an important role in the detection of internal inconsistencies. The main effect of grade implicates developmental differences in these skills, while the effect of reading level probably reflects the effect of general intelligence.

Discussion

The present study has provided a number of important insights into the ways children evaluate their understanding as they read. Although previous studies have provided some evidence regarding children's use of specific criteria, none have focused on multiple standards. Moreover, the results have typically been interpreted as though comprehension monitoring were a global entity, something at which a child is either effective or ineffective. This simplistic conception of comprehension monitoring must be abandoned if we are to effect any changes in children's ability to decide for themselves whether or not they understand.

The present study shows quite clearly that there are children who in fact are limited in their evaluation skills. These limitations are reflected in several different dimensions of the data. Consider, for example, children's identification of the intentionally-introduced problems. Poorer readers were less successful at identifying the problems than better readers, consistent with several other studies (e.g., Garner & Kraus, 1981-82; Paris & Myers, 1981). Additionally, younger children were less successful than older children, but the age-related change was found only among the better readers. The fact that the older poorer readers identified no more problems than the younger poorer readers has important implications. One interpretation of the findding is not that the older students fail to improve in their ability to evaluate their understanding, as the results might suggest, but rather that they exhibit an increasing lack of confidence or incentive to do so. Some support for this hypothesis is provided by the finding that all of the poorer readers benefitted from instructions specifying the types of problems they

should seek, even though the benefit was not as great as for the better readers. But a more telling argument is that a substantial proportion of the older poorer readers never used the lexical standard of evaluation, even though this was the standard most likely to be adopted by the majority of the children. Whether this apparent reluctance to admit word comprehension failures characterizes the way the students typically respond internally to the demands of reading or whether it only occurs externally in interactions with other people is an important empirical question.

Differences among the children were also apparent in the size and composition of their repertoire of standards. Better readers used more different standards than poorer readers, regardless of whether they were instructed as to the kinds of standards they should use. This suggests that they routinely evaluate their understanding with respect to more different criteria than the poorer readers. Additionally, although fourth graders who received specific instructions did not differ from sixth graders in the number of standards used, fourth graders who were left to adopt whatever criteria they chose had a more limited repertoire than their sixth grade counterparts. Note in particular that the fourth grade poorer readers used an average of only 1.02 different standards under general instructions. In other words, they tended to rely exclusively on a single standard.

Among those children who used a single standard, it was virtually always the lexical standard that was adopted. This finding is consistent with the often-reported emphasis on word understanding among younger and poorer readers (e.g., Garner, 1981; Myers & Paris, 1978). Even when specifically told to use other standards, close to 25% of the younger readers did not. Recall that all

children did in fact know that the passages contained problems that were defined as things that might confuse people or that they might have trouble understanding. The children did not seem to realize that there were other possible sources of comprehension difficulty. It seems, then, that the way many children typically decide whether or not they understand is by checking to make sure individual word meanings are known.

Despite this higher incidence of reliance on the lexical standard among younger and poorer readers, there were also many children in these same groups who never used the standard. As noted earlier, this pattern was most pronounced among the sixth grade poorer readers. Quite clearly, there are individual differences in the standards used by less effective readers. Failures to question word understanding at all may be just as detrimental as failures to consider anything but word understanding. Although a number of good readers also never used the lexical standard, responses to the follow-up questions revealed a different pattern of dealing with the nonsense words. The better readers tended to have figured out during reading a plausible meaning for the nonsense words on the basis of surrounding context, while the poorer readers, even at the time of questioning, had difficulty coming up with a plausible meaning.

Children's use of the external consistency standard also varied with age and reading proficiency. Exclusive use of the standard was rare, but it was somewhat more common among poorer readers. Among the children who never used the standard, there were more fourth graders than sixth graders. Although poorer readers were no less likely to use the standard than better readers, they tended to use it more frequently and less effectively. Note in

particular that the younger poorer readers who received specific instructions challenged the truth of 8.58 propositions on the average yet the mean number of prior knowlege problems they identifed was only 1.11. Morever, many poorer readers who used the standard failed to identify any of the embedded problems.

The internal consistency standard was present in only 55% of the subjects' repertoires; in other words 45% of the chidlren never questioned the consistency of any of the ideas within the passages. More younger and poorer readers fell into this grouping, as did those receiving general instructions. The fact that so many children never used the standard at all accounts for the low detection of internal inconsistencies in the present study and it also suggests an explanation for the poor inconsistency detection reported in other studies (e.g., Garner, 1981; Garner & Kraus, 1981-82; Markman, 1979). Many children do not think to evaluate their understanding with respect to internal consistency, and even when they are instructed to adopt such a standard, they still do not use it frequently, let alone effectively. Consider, for example, the fourth grade poorer readers: on the average, they used the internal consistency standard less than once throughout the entire session. Evaluation of internal consistency requires careful processing of the text. In contrast to the external standard, its use cannot be "faked," as evidenced by the fact that only two of the 108 subjects used the standard at least once but did not identify any inconsistencies.

Although the primary focus of the study was on children's use of three specific standards for evaluating their understanding, the study provides some evidence of the use of other standards as well. The most frequently used

non-target standard was informational completeness and clarity. A total of 14 children used the standard at least once, 11 of whom were more proficient readers. Specificity of the instructions did not influence its use. Although the proportion of children using the standard was small, the reading proficiency difference suggests that better readers are more likely to spontaneously consider whether the text contains sufficient information to enable them to grasp the main ideas. Comments indicating that other standards had been applied were more infrequent, probably because the passages had been screened by adult readers for the presence of other problems. Four children used the structural cohesiveness standard, one the propositional cohesiveness standard, and five the syntactic. Additional research is needed to examine more directly the extent to which children evaluate their understanding with respect to these other standards [illegible]. If we wish to improve readers' abilities to decide for themselves when they understand and when they do not, we must have a thorough understanding of the kinds of criteria they do and do not use.

Reference Notes

1. Baker, L. Children's effective use of multiple standards for evaluating their comprehension. Unpublished manuscript. University of Maryland Baltimore County, 1983.

References

Baker, L. How do we know when we don't understand? Standards for evaluating text comprehension. In D.L. Forrest, G.E. Mackinnon, & T.G. Waller (Eds.), Metacognition, cognition, and human performance. New York: Academic Press, in press.

Garner, R. Monitoring of understanding: An investigation of good and poor readers' awareness of induced miscomprehension of text. Journal of Reading Behavior, 1980, 12, 55-64.

Garner, R. Monitoring of passage inconsistency among poor comprehenders: a preliminary test of the "piecemeal processing" explanation. Journal of Educational Research, 1981, 74, 159-162.

Garner, R., & Kraus, C. Monitoring of understanding among seventh graders: an investigation of good comprehender-poor comprehender differences in knowing and regulating reading behaviors. Educational Research Quarterly, 1982, 6, 5-12.

Garner, R., & Taylor, N. Monitoring of understanding: an investigation of attentional assistance needs at different grade and reading proficiency levels, Reading Psychology, 1982, 3, 1-6.

Grice, H.P. Logic and conversation. In P. Cole & J.L. Morgan (Eds.) Syntax and semantics (Vol 7): Speech Acts. New York: Academic Press, 1975.

Harris, P.L., Kruithof, A., Terwogt, M., & Visser, T. Children's detection and awareness of textual anomaly. Journal of Experimental Child Psychology, 1981, 31, 212-230.

Harris, A. J., & Sipay, E. R. How to increase reading ability. New York: Longman, 1980.

Markman, E.M. Realizing you don't understand: Elementary school children's awareness of inconsistencies. Child Development, 1979, 50, 643-655.

Markman, E. M. Comprehension monitoring. In W. P. Dickson (Ed.), Children's oral communication skills. New York: Academic Press, 1981.

Markman, E.M., & Gorin, L. Children's ability to adjust their standards for evaluating comprehension. Journal of Educational Psychology, 1981, 73, 320-325.

Paris, S.G., & Myers, M. Comprehension monitoring, memory, and study strategies of good and poor readers. Journal of Reading Behavior, 1981, 13, 5-22.

Spiro, R., Bruce, B. C., & Brewer, W. F. (Eds.) Theoretical issues in reading comprehension. Hillsdale, N.J.: Erlbaum, 1980.

Footnotes

The research reported in this paper was supported in part by the National Institute of Education under Grant NIE-G-81-0100. I am grateful to the students and staff of the Baltimore County Public Schools for their cooperation and participation. I thank Susan Sonnenschein for her comments on the manuscript. Address reprint requests to the author at the Department of Psychology, UMBC, Catonsville, Maryland 21228.

1. Since there were only three examples of each problem type, it is important to know the extent of variablity among items, even though an analysis treating items as a random effect is inappropriate. Therefore, the frequency with which each specific problem was identified was compared to the frequencies for other problems of the same type and different types. Despite some variability, all of the nonsense word problems were identified more often than any of the prior knowledge or internal inconsistency problems ($p = .47, .59, .60$). All of the prior knowledge problems had similar identification rates ($p = .36, .37, .43$) as did the internal inconsistencies (.33, .33, .43). Additionally, within conditions, there were no specific items which had detection rates grossly different from the overall pattern.

2. Since the means were based only on missed problems, a child who did not miss any problems would have no score. Seven children fell into this category. In addition, the taped protocols for three of the children were incomplete and so their scores could not be calculated. This reduction in sample size accounts for the reduced degrees of freedom.

Table 1

Mean Number of Intentionally-Introduced Problems Identified

			Type of Problem		
			Nonsense Word	Prior Knowledge Violation	Internal Inconsistency
Grade	Reading Proficiency	Instruction			
Fourth	Better	Specific	2.13	1.54	1.57
		General	.98	.44	.62
	Poorer	Specific	1.14	1.11	.90
		General	1.25	.29	.14
Sixth	Better	Specific	2.56	2.25	2.56
		General	1.98	1.54	1.42
	Poorer	Specific	1.10	1.00	1.20
		General	1.09	.96	.34

Note--Maximum in each cell is 3.

Table 2

Mean Number of Times Standards Were Applied

Throughout Testing Session

		Type of Standard			
Grade	Reading Proficiency	Instructions	Lexical	External Consistency	Internal Consistency
Fourth	Better	Specific	3.00	4.20	1.93
		General	1.69	1.25	1.19
	Poorer	Specific	4.67	8.56	.83
		General	3.45	1.18	.18
Sixth	Better	Specific	2.94	4.94	3.13
		General	2.44	3.25	1.75
	Poorer	Specific	1.60	4.70	1.40
		General	2.17	3.42	.33

Table 3

Proportion of Children Who Never Used

A Particular Standard of Evaluation

Grade	Reading Proficiency	Instructions	Type of Standard: Lexical	External Consistency	Internal Consistency
Fourth	Better	Specific	.00	.20	.27
		General	.56	.69	.56
	Poorer	Specific	.06	.25	.50
		General	.18	.62	.91
Sixth	Better	Specific	.00	.00	.00
		General	.13	.25	.31
	Poorer	Specific	.30	.10	.40
		General	.33	.33	.67

Table 4

Proportion of Children Who Only Used

One Particular Standard of Evaluation

			Type of Standard		
Grade	Reading Proficiency	Instructions	Lexical	External Consistency	Internal Consistency
Fourth	Better	Specific	.20	.00	.00
		General	.51	.00	.06
	Poorer	Specific	.25	.08	.00
		General	.73	.09	.00
Sixth	Better	Specific	.00	.00	.00
		General	.19	.00	.00
	Poorer	Specific	.10	.20	.00
		General	.42	.08	.00

Table 5

Mean Number of Different Standards Used Throughout Testing Session

Reading Proficiency	Instructions	Fourth	Sixth
Better	Specific	2.68	3.13
	General	1.67	2.62
Poorer	Specific	2.46	2.40
	General	1.02	1.66