1-1-2009

Using Semantic Ambiguity Instruction to Improve Third Graders' Metalinguistic Awareness and Reading Comprehension: An Experimental Study

Marcy Zipke

Linnea C. Ehri

Helen Smith Cairns

Using Semantic Ambiguity Instruction to Improve Third Graders' Metalinguistic Awareness and Reading Comprehension: An Experimental Study

Marcy Zipke
Providence College, Rhode Island, USA

Linnea C. Ehri, Helen Smith Cairns
Graduate Center of the City University of New York, New York, USA

ABSTRACT

An experiment examined whether metalinguistic awareness involving the detection of semantic ambiguity can be taught and whether this instruction improves students' reading comprehension. Lower socioeconomic status third graders (*M* age = 8 years, 7 months) from a variety of cultural backgrounds (*N* = 46) were randomly assigned to treatment and control groups. Those receiving metalinguistic ambiguity instruction learned to analyze multiple meanings of words and sentences in isolation, in riddles, and in text taken from the Amelia Bedelia series (Parish, 1979, 1988). The control group received a book-reading and discussion treatment to provide special attention and to rule out Hawthorne effects. Results showed that metalinguistic ambiguity instruction was effective in teaching students to identify multiple meanings of homonyms and ambiguous sentences and to detect inconsistencies in text. Moreover, this training enhanced students' reading comprehension on a paragraph-completion task but not on a multiple-choice passage-recall task, possibly because the two tests differ in the array of linguistic or cognitive correlates influencing performance. Comprehension monitoring was not found to mediate the relationship between ambiguity instruction and reading comprehension. Results carry implications for the use of language-based methods to improve reading comprehension in the classroom.

Metalinguistic awareness (MA) is the ability to focus on and manipulate the formal properties of language—specifically, the ability to analyze, think about, talk about, or play with language as an object separate from its meaning in or out of context (Roth, Speece, Cooper, De La Paz, 1996). Various types of MA have been distinguished, including phonological awareness, syntactic awareness, and morphological awareness. MA is regarded as having special importance for helping students learn to decode words and to comprehend text. As a result of much research, phonemic awareness has been established as an essential contributor in learning to read (Ehri, Nunes, Stahl, & Willows, 2001). Other types of MA, however, have not attained this status, although there has been substantial research indicating the benefits of instruction in morphological awareness (Carlisle, 1995) and syntactic awareness (Gaux & Gombert, 1999; Nation & Snowling, 2000; Tunmer, Nesdale, & Wright, 1987). The purpose of the present study was to examine whether reading comprehension is benefited by instruction in another type of MA, namely semantic ambiguity detection, which requires recognizing when one linguistic form, either a word or a sentence, has two different meanings.

Various theories have been proposed to explain the contribution of MA to reading. In the model proposed by Tunmer and Bowey (1984), young children acquire implicit knowledge of the structure of language in the context of learning to comprehend and communicate shared meanings with others. Subsequently, when they enter school and begin learning to read, their focus on meaning is set aside and they shift attention to the form of language, including phonemic and lexical units and their relationships to letters and spelling units. It is not until the decoding skill is securely in place that children move away from treating words as isolated units and pay attention to their role in sentences. Tunmer and Bowey (1984) likened this to "put[ting] humpty-dumpty back together again" (p. 163). For some students, however, the ability to refocus attention from form to meaning in written language does not come easily. Tunmer and Bowey have suggested that what these students need is MA—specifically syntactic awareness—to help them integrate word units to form meaningful sentences and texts and to enable them to monitor their comprehension. Greater facility with self-monitoring in turn serves to boost their reading comprehension.

Cairns, Waltzman, and Schlisselberg (2004) investigated the development of the metalinguistic skill of ambiguity detection and its relation to reading in the early school years. Two of the basic hallmarks of metalinguistic skill are the ability to attend simultaneously to the form and content of language and the ability to think and talk about language as an object rather than simply as a means of communication (Hakes, 1980; van Kleeck & Reddick, 1982). People who can perceive and report that a sentence such as "The man's nails were sharp" has two meanings but only one form demonstrates both of these metalinguistic abilities. The perception that a single sentence form can have dual content demonstrates the first hallmark, and the ability to talk about the sentence and its two meanings demonstrates the second.

Cairns et al. (2004) tested children on their ability to report the ambiguity of sentences whose ambiguity rested on the dual meanings of a homonym, like the "nails" sentence above (lexically ambiguous sentences), and also on their ability to detect structurally ambiguous sentences. The latter are sentences without ambiguous words, whose dual meaning derives from the fact that they have two possible structural organizations. For instance, the sentence "The girl tickled the baby with the stuffed bear" can mean that either the girl or the baby had the bear. The former meaning is reflected in an underlying structure in which the "bear" is used as an instrument by "the girl," the latter in a structure in which the "bear" is possessed by "the baby" (Cairns, 1999). Cairns et al. (2004) demonstrated that 4- and 5-year-old children failed to report both kinds of ambiguity. First graders could not perceive and report structural ambiguities, but they were able to detect some lexical ambiguities. This was early in the first grade, when the children were prereaders; regression analyses showed that their lexical ambiguity-detection scores in first grade accounted for more than half the variance in their second-grade reading scores. By second grade, the same children demonstrated an ability to detect structural ambiguities, and both detection scores were significant predictors of their third-grade reading scores.

Cairns et al. (2004) suggested two explanations for the relationship between ambiguity-detection skill and reading ability. One relates to the metalinguistic skill required to perform the detection tasks, the other to the operation of psycholinguistic processes. The metalinguistic aspect of ambiguity detection depends upon the child having available two meanings of the ambiguous sentence. In the case of lexical ambiguities, the two sentence meanings each depend upon a different meaning of the ambiguous lexical item. In the case of structural ambiguity, the two sentence meanings are determined by distinct structural representations of the sentence. In both cases, the two representations are derived by the same psycholinguistic processes that children (and adults) use to understand all spoken sentences.

Decades of research in sentence processing have resulted in a good understanding among psycholinguists of exactly how these processes work (Cairns, 1999). To understand a sentence, both children and adults must retrieve the words of the sentence from their internalized lexicons and construct a syntactic representation of the sentence. Works by Swinney (1979), Swinney and Prather (1989), and Love, Swinney, Bagdasaryan, and Prather (1999) have demonstrated that when readers or listeners process a sentence containing an ambiguous word, both meanings of the ambiguous lexical item are retrieved from the internal lexicon and a second operation rapidly selects the appropriate, contextually relevant meaning. (Meaning retrieval is independent of prior context, but meaning selection is affected by context.) Structural analysis of the sentence is determined by the syntax of the hearer's/reader's grammar, as well as other processing strategies and preferences. Thus, to conduct a dual analysis of a lexically ambiguous sentence, the hearer/reader must perform two operations: a first in which one meaning is constructed by incorporating the preferred meaning of the homonym, then a second in which reprocessing creates a sentence incorporating the second meaning. Note that the ability to simply identify a homonym is not sufficient to perceive the ambiguity of a lexically ambiguous sentence (Shakibai, 2007). Similarly, to conduct a dual analysis of a structurally ambiguous sentence, the hearer/reader must create two different syntactic forms of the sentence. It is important to realize that the sentence-processing operations described above are carried out rapidly and unconsciously, whereas the metalinguistic operations that come into play during ambiguity detection are not only conscious but also require deliberate contemplation on the part of the person who is making the judgment.

One limitation of the Cairns et al. (2004) study of ambiguity detection with children is that the findings were correlational. To determine whether ambiguity detection directly contributes to reading ability, an experiment is required. The purpose of the present study was to examine whether teaching students to reprocess and restructure sentences to recognize ambiguity would improve their reading comprehension compared with a control group that did not receive this same instruction.

One type of ambiguity detection familiar to children consists of understanding riddles. Riddles are fun and their texts are short, making them particularly appropriate for assessing and teaching ambiguity detection in younger students. Various researchers have studied riddle comprehension in children (Fowles & Glanz, 1977; Hirsh-Pasek, Gleitman, & Gleitman, 1978; Mahony & Mann, 1992, 1998; Shultz, 1974; Yalisove, 1978). However, Yuill (1996, 1998) was the first to examine experimentally whether using riddles to teach children about ambiguity would benefit their reading comprehension.

Yuill (1998) based her studies on Tunmer and Bowey's (1984) model, in which different types of metalinguistic skills are predicted to have an impact on different aspects of reading ability, with some contributing to decoding and others to comprehension. In support of this, Yuill (1996, 1998) found that the ability to solve morpho-phonological riddles was positively and significantly correlated with word-reading accuracy, whereas the ability to solve riddles at the lexical and syntactic levels correlated significantly with reading comprehension. This led Yuill (1998) to study whether metalinguistic training in riddles and ambiguities improved the reading comprehension of 7- and 8-year-olds. Eighteen children whose reading comprehension lagged significantly behind their decoding skill were matched on decoding skill, age, and vocabulary with 18 good comprehenders, and members of matched pairs were randomly assigned to experimental and control groups. The experimental group received seven weekly treatment sessions. Children learned about the double meaning of words in isolation and in sentences. They made up jokes using word compounds. They played a communication game involving ambiguous messages and a clue-construction game to make decisions about word meanings. They learned riddles that involved a contrast between the mention and use of a word, for example, "What word is loud, even when you say it soft? Loud." They read ambiguous stories. The control group read silly stories and played phonemic awareness games. Results revealed that although both groups showed improved reading comprehension following training, the children given ambiguity training significantly outperformed controls by about six months in comprehension age.

Yuill (1998) interpreted her results as indicating that teaching struggling readers to be flexible with words and to attend to syntax is a form of MA that is effective in boosting reading comprehension. However, an alternative explanation is that MA enhanced students' comprehension monitoring (CM) skill, which acted as an intervening variable to improving their reading comprehension. According to Tunmer and Bowey (1984), although poor comprehenders have implicit syntactic knowledge, they fail to use this knowledge to monitor their comprehension and repair comprehension failures when reading text. Effective CM requires detecting when the text does not make sense and then taking steps to resolve the problem and restore meaning by reanalyzing structural relations within sentences. Ambiguity training may have given students sufficient access to their implicit syntactic knowledge so that they could revisit and reorganize syntactic relations within sentences, hence improving their metacognitive control over their reading of texts.

The aim of the current study was to conduct an experiment to confirm and extend the findings of Yuill (1998) and Cairns et al. (2004). We developed a series of ambiguity-detection training procedures that included activities similar to those described by Yuill. (Note that the terms *instruction* and *training* are used synonymously in this article.) Students were taught to reprocess ambiguous words, sentences, and riddles until the second meanings became evident. Prior to and following training, students' abilities to detect lexical and syntactic ambiguities with words, sentences, and riddles were measured to assess whether the instruction was effective. Students' reading comprehension was tested to determine whether training made any contribution. Their CM skill was assessed to examine whether this served a mediating function linking MA to reading comprehension.

Students' MA was operationalized as ambiguity detection at the word and sentence level. The word-level task required students to explain two conceptually different meanings of individual words spoken by the experimenter. The sentence-level task required the same response to sentences. The riddle task presented children with questions in standard riddle form and required them to select which of two punch lines would form a riddle.

Students' ability to monitor their oral and written language was operationalized with three tasks that required different types of CM. Effective comprehension requires accurate word and sentence processing, so a varied set of monitoring tasks was included. To assess monitoring at the word level, students were asked to read aloud from sentences that contained heteronyms, or words that are spelled the same but have different pronunciations (e.g., *bow* and arrow versus taking a *bow* to an audience). The students' scores indicated whether they varied the pronunciations of the heteronyms to reflect their different meanings. Because the heteronym sentences were presented in isolation, another measure of word-level monitoring was included, one involving the reading of words in text. Students read from a series of graded texts and were scored according to whether or not they self-corrected miscued words as they read. Finally, students' CM was measured on the story level. Like Baker and Zimlin (1989) and Hacker (1997), students' abilities to coordinate the meanings of propositions in a text were measured with an error-detection paradigm. To determine whether students detected anomalous information embedded in a story, this task was administered as a think aloud. Although three tasks were included, the error-detection task was regarded as having the strongest face validity as a measure of CM.

Two hypotheses were tested: (1) MA that involves the detection of semantic ambiguities can be taught to third graders, and (2) this instruction will improve their CM skill and their reading comprehension. To assess the effectiveness of instruction, pretests and posttests were given. In addition, a noninstructed control group was included. Control students were drawn from classrooms receiving the same literacy instruction as the treatment group. They were made to feel like participants receiving a special treatment that involved meeting in small groups to read and discuss stories.

The present study was expected to yield important findings. Teachers and parents struggle to find more effective and innovative ways to improve children's ability to read and comprehend text, especially those at a lower socioeconomic status, like those in the current study. Research has shown that instruction in decoding, vocabulary, fluency, background knowledge, and cognitive and metacognitive strategies all contribute to reading achievement (Pressley, 2000). Fewer studies have investigated the importance of MA involving ambiguity detection. However, based on the research reviewed here, there is good reason to believe that enhancing students' awareness of semantic ambiguities can aid in their CM and their reading comprehension.

Methods

Participants

Participants were third graders recruited from two public schools in the same geographic area of a large, northeastern U.S. city in the winter. The population of the schools was from a lower socioeconomic status and included students from a variety of ethnic backgrounds, with 63% of the students qualifying for a free or reduced-cost lunch. Approximately 5% of the students spoke English as a second language. Participation in this study was limited to fluent English speakers and to those who did not have individualized education programs that qualified them for special education.

All of the children who returned parental consent forms were pretested for reading ability and receptive vocabulary knowledge. Reading-comprehension scores were used to form matched pairs of students. Potential pairs were first matched as closely as possible on their scores on the Woodcock Reading Mastery Test-Revised (WRMT-R) Reading Comprehension subtest (Woodcock, 1987), then on their scores on the Gates-MacGinitie Reading Test, Fourth Edition, Level 3 (GMRT4; MacGinitie, MacGinitie, Maria, & Dreyer, 2000). Matches were made from within the same school but not necessarily from the same classroom. Members of pairs were randomly assigned to the experimental and control groups ($n = 23$ per group).

Materials and Procedures

To ensure that all participants spoke English fluently and had adequate decoding skills for the training, several pretests were administered, including the Word Identification and Word Attack subtests of the WRMT-R (Woodcock, 1987) and the third edition of the Peabody Picture Vocabulary Test (Dunn & Dunn, 1997). Each potential participant was removed from the classroom individually and briefed about participation in the study. They then completed the vocabulary test followed by the Word Identification and Word Attack subtests. One child with very low reading scores was excluded from the study.

Eight tasks were administered as pretests and then repeated as posttests following training. In all cases, the posttest was identical to the pretest except for the items. There were three measures of MA assessing ambiguity detection, three measures of CM, and two measures of reading comprehension. Tests 1 through 7 were administered in a single session of about 45 minutes, whereas Test 8 required a separate session lasting 45 minutes. All pretests were given prior to training, and posttests were administered from three to seven days after training ended. All of the tests were individually administered, except for the GMRT4 Reading Comprehension subtest, which was administered to small groups. The tests were administered in the following order.

Pretests and Posttests

Homonym Definition

To assess ambiguity detection at the word level, participants listened to 10 homonyms and described as many meanings for each word as possible. After each definition response, the participant was asked, "Does ___ mean anything else?" Students often responded with additional instances of a single meaning (e.g., *sink*: "A place where you wash your hands" and "A place where you wash dishes") rather than a different meaning (e.g., *sink* as the verb, meaning to fall down in water). The number of words given two distinct definitions was considered the student's score. For example, if a student gave uses of a word that showed two different parts of speech or if the student was able to use the word in two different contexts that made the separate meanings clear, the item was scored correct. Scoring for this task was carried out blind to condition. The Alpha reliability was .82. (See Appendix for items.)

Ambiguous-Sentence Detection

To assess ambiguity detection at the sentence level, participants listened to eight ambiguous sentences, one at a time, and explained as many alternative meanings as possible (as in Cairns et al., 2004). Prior to testing, participants heard the sentence, "The chicken is ready to eat" and were shown two pictures: In Figure 1, the chicken is hungry and ready to eat its dinner, and in Figure 2, the chicken is cooked and ready for other people to eat. A second practice item, "They talked about the problem with the teacher," was also presented and the two meanings discussed. Then the eight test sentences were given. Participants were told to explain one meaning of the sentence and then to think of another meaning. What was scored was the number of instances in which the student correctly explained two distinctive meanings. Scoring for this task was carried out blind to condition. Alpha reliability was .83. (See Appendix for items.)

Riddle Resolution

In this ambiguity-detection task adapted from one used by Zipke (2007), the experimenter read aloud five questions, each followed by two punch lines, and the child selected the punch line that created a riddle. First, the concept of a riddle was explained: "A riddle is a puzzling question that ends with an answer that surprises you and usually makes you laugh. The question and answer make a riddle when the same words have two different meanings." Children were further instructed that only one of the punch lines contained an ambiguity that gave the riddle multiple meanings, and their job was to choose the punch line that made the question into a riddle. For example, "Why did the skeleton go to the movies by himself?" "He had no body to go with him" or "He was lonely." In this example, "He had no body to go with him" is the correct choice because there are two possible meanings for "no body": It could mean that the skeleton did not know another person to accompany him, or it could mean he had no physical body. The number of correct punch lines was scored. Because the test consisted of 5 two-choice items, chance-level performance was 2.5 items correct. Alpha reliability was .65. (See Appendix for items.) The reliability was lower than we would have liked but was considered acceptable for the purposes of the study. The fewer test items and the influence of guessing might have contributed to the lower reliability.

Heteronym Pronunciation

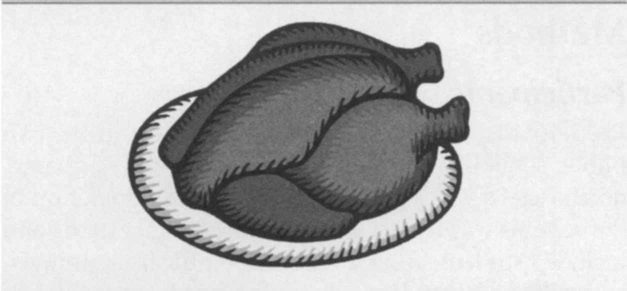
To assess CM at the word level, participants read aloud 10 sentences, each containing two heteronyms, that is, words that are spelled the same but pronounced differently (e.g., "I lowered my bow and arrow and took a bow"). Participants were told only to read each sentence aloud. Scored was the number of sentences read with the heteronyms pronounced correctly, either on first pass or self-corrected. Alpha reliability was .81.

Miscue Self-Correction

To assess CM during text reading, participants read aloud from a series of graded paragraphs that increased in difficulty until their errors exceeded 10% of the text. Two scores were calculated on passages that were read at or above 90% accuracy: The number of miscues (i.e., word substitutions) that were produced and the proportion that were self-corrected.

Figure 1. "The Chicken Is Ready to Eat" Definition 1

Figure 2. "The Chicken Is Ready to Eat" Definition 2



Anomaly Detection

To assess CM on the story level, participants read aloud from familiar stories (i.e., The Three Little Pigs at pretest and The Three Bears at posttest) containing micro- and macrostructure anomalies. Baker and Zimlin (1989) conceptualized microstructure standards as those requiring local attention to words, sentences, structural cohesiveness, and internal consistency (e.g., "'I'm a big bad wolf!' said the hungry bear") and macrostructure standards as those requiring external knowledge of the world (e.g., "Winter was coming soon, when the weather would be very hot"). Children were instructed to read aloud and, stopping at every star they saw inserted in the story, to, "Tell me what you are thinking. For example, if you see something you think is wrong, or doesn't make sense, or if what you are reading reminds you of something else, or if you think you know what will happen next, tell me about it." Responses were taped, transcribed, and scored by two independent raters. The number of anomalies mentioned in children's comments was counted. Alpha reliability was .92.

WRMT-R Reading Comprehension

The Reading Comprehension subtest from the WRMT-R (Woodcock, 1987, 1998) was administered individually. Form G was given as the pretest and Form H as the posttest. Participants silently read several short passages, each containing a blank space. They responded by filling in the blank to complete the meaning of the passage. According to the publisher's manual, the split-half reliability is .92.

GMRT4 Reading Comprehension

The GMRT4 was administered, Form S as a pretest and Form T as a posttest, during a separate session lasting 45 minutes. This was a timed test in which participants read 11 passages silently, each approximately three paragraphs in length, and answered multiple-choice comprehension questions. The test was group-administered to three or four students at a time. Raw scores were converted to extended scale scores based on test norms. According to the publisher's manual, the Kuder-Richardson Formula 20 reliability is .91.

Training in MA Involving Ambiguity Detection

Training was conducted individually, in sessions lasting about 45 minutes, once per week over the course of four weeks. One investigator trained and tested all participants. She removed them from the classroom at times approved by the classroom teacher and took them to the school library or resource room. Participants received training on words and sentences with multiple meanings. Instruction began at the word level and progressed to the reading of authentic texts. Active learning methods, which required participants to manipulate objects and produce responses on their own, were employed using materials and procedures typically found in public school classrooms. New concepts were presented and modeled, then turned over to students for guided practice. Participants performed the following procedures:

Session 1: Multiple Word Meanings

Students were taught that words can have more than one meaning. The instructor and participant brainstormed and discussed words that they knew to have more than one meaning. Then the student was given a tub containing 40 words, including nouns, verbs, adjectives, and adverbs, printed on tiles and was challenged to find the homonyms (22 of the 40 words were homonyms).

Two primary strategies were taught for identifying homonyms. The first strategy was to consider whether a word, such as *watch,* could fit another form class (e.g., if it's a thing, can it also be something you do?). The second strategy was to generate synonyms to uncover potentially different definitions and then to assess whether the synonyms were truly different (e.g., for the word *ball*, baseballs and basketballs are both still sports equipment so they are not really different meanings of *ball*).

Session 2: Multiple Sentence Meanings

Instruction consisted of two parts. First, the participants explored seven ambiguous sentences, each accompanied by two illustrations representing different meanings of that sentence. For example, children heard the sentence "The dog chased the man on a bike" and were shown one illustration featuring a dog running after a bicycle with a man atop it and a second illustration featuring a man running away from a bicycle-riding dog. Their task was to explain how each picture illustrated the sentence. The experimenter provided scaffolded support to help children complete the task.

Next, the children were challenged to represent the two meanings of eight ambiguous sentences by using Colorforms (a type of reusable, manipulable sticker) and attaching them to pictures displaying a relevant context. For example, in response to the sentence, "The ball was found by the kitten," children were given relevant Colorforms (i.e., ball, kitten) and expected to attach them to the picture in two different ways to represent the two meanings of the sentence: (1) by hiding the ball near the kitten and adding another character who finds the ball and (2) by placing the ball anywhere on the scene and then having the kitten find the ball. When necessary, student efforts were scaffolded by analyzing different sentences with the same form (i.e., "The baseball bat was found by the monkey" or "The seashell was found by the penguin").

After the first sentence was practiced, the instructor introduced a poster board chart labeled with a large *W* atop each of three columns. She explained that the three *W*s stand for *who does what to whom*. As they progressed through the sentences, the experimenter wrote down answers for each interpretation of the sentences in the appropriate columns.

Session 3: Riddles

The instructor introduced the concept of lexical riddles first: "Why do spiders like baseball? They're good at catching flies." She explained that questions and punch line answers are riddles when they contain words with two different meanings. Then she introduced structural riddles: "Where can you see a man eating fish? A seafood restaurant." She used the 3W chart and manipulables (if possible) to explain the answers. The participants were asked to volunteer any riddles they knew.

Loosely following the procedures laid out in Bernstein (1979) for writing riddles with children, the instructor and student identified a topic of particular interest to the student. They brainstormed to create a list of words pertaining to that topic and then looked for homonyms in the list. They made up questions that seemed to involve the primary meaning but were really about the alternative meaning. For example, if the topic was Harry Potter, a common word was *spell*, so one student wrote the riddle, "Why does Harry Potter go to school? To learn how to *spell* better." A popular topic was baseball, yielding homonyms such as *batter, shortstop, diamond, plate*. An example of one of the participant's baseball riddles was, "Why were the baseball players on strike? They couldn't hit the ball!" (For more information, see Zipke, 2008.) Children were scaffolded in their attempts to use the homonyms to produce riddles that were written down in their own riddle book.

Session 4: Text Reading

In this final session, the participants graduated to book reading. First the children read *Amelia Bedelia and the Surprise Shower* (Parish, 1979). If students had trouble decoding a word, help was provided. In the Amelia Bedelia series, Amelia frequently misunderstands directions. She does things like "trim" the steak with ribbons and lace or "make a jelly roll" by prodding jelly along the floor. In this way, the books include both lexical and structural ambiguities lodged in individual sentences.

The children were instructed to stop at every sentence with more than one meaning and explain how Amelia understood the sentence, as well as how the sentence was meant to be understood. Finally, they explained how they knew which was the intended meaning.

Next, the children read *Amelia Bedelia's Family Album* (Parish, 1988). On the first of several alternating pages, Amelia introduced her family members one by one, with their names and professions and their employers' conventional interpretations of the professions. The next page displayed Amelia's alternative interpretation. For example, a "boxer" to Amelia was someone who put things in boxes. After the first few examples, the instructor covered Amelia's interpretations in the book and required students to speculate on what Amelia would say. Finally, the participants added their own entries into Amelia's family album with professions brainstormed by the participant and experimenter.

Control-Group Activities: Book Reading and Discussion

The purpose of the control group was not to compare metalinguistic training to an alternative form of instruction; it was to provide a noninstructional treatment baseline for assessing effects of metalinguistic training. To eliminate Hawthorne effects, the control group was led to believe that it was actively participating in the experiment and was receiving special treatment. Classroom teachers reinforced the belief that all students were receiving valuable experiences that should benefit their reading.

Control-group participants were removed from their classrooms and met with the experimenter once per week for four weeks but for shorter periods of time (often 10–15 minutes) and in groups of two or three to read aloud and discuss the book *Mouse Soup* (Lobel, 1977). Discussion activities included identifying components of the book (i.e., illustrator, title page, table of contents, etc.); making story maps; and talking about plot, point of view, setting, and imagery. However, semantic ambiguities were never discussed. Participants were told that the investigator was trying to figure out the best way to teach kids to read. All of the students believed they were receiving special reading instruction.

Design and Statistical Analyses

A pretest/posttest experimental design with random assignment to treatment and control conditions was used. The experimental group received ambiguity-detection training while the control group received special attention in the form of storybook reading and discussion but no training in MA. Students were matched on reading-comprehension pretests to randomly assign pair members to conditions. Pretests of the two groups were compared with matched-pair t tests to verify that the treatment and control groups did not differ significantly prior to training. Analyses of variance (ANOVAs) were used to assess the effectiveness of metalinguistic training on outcome measures. The independent variables were treatment (ambiguity training vs. control) and time of test (pretest vs. posttest). The treatment variable was analyzed statistically as two independent groups, whereas time of test was a repeated measure. Significant interactions were sought as evidence that the treatment group made greater gains from pretest to posttest than did the control group. Correlations were examined to study relationships of special interest among pretest, training, and posttest measures.

Hypotheses were tested at an α level of .05, except when correlations of special interest were analyzed. In this case, hypotheses involving the whole sample ($N = 46$) were tested at an α level of .01. Because several correlations were examined, this level limited the possibility of Type 1 errors. However, in the case of hypotheses involving correlations tested on the group receiving ambiguity instruction, an α level of .05 was adopted because the small sample size (i.e., $n = 23$) limited the power to detect significant relationships. Results in this case were considered suggestive.

Results

Characteristics of Participants

Tables 1 and 2 present mean performance and test statistics comparing the treatment and control groups on pretests. *T* tests confirmed that the groups did not differ significantly in age or on any of the reading or vocabulary tests (see Table 1). Also, *t* tests verified that the groups did not differ significantly on the MA and the CM tasks (see Table 2).

As is evident in Table 1, the third graders' mean grade-equivalent scores on subtests of the WRMT-R placed them at late second grade on the Word Attack test, late third grade on the Word Identification subtest, and early third grade on the Reading Comprehension subtest. Thus, they were reading more or less as expected for their grade level. Their mean standard score on the Peabody Picture Vocabulary Test indicated an average vocabulary level.

Although we employed no objective measure of this, it was clear from observation that the students enjoyed and were engaged with the treatment activities

Table 1. Characteristics and Mean Performance of the Metalinguistic Treatment and Control Groups on the Language and Reading Pretests

Pretest	Treatment group	Control group	$t(22)$ value
Age in months	103.22 (3.42)	104.65 (4.01)	−1.31
Gender			
Boys	11	14	
Girls	12	9	
Ethnicity			
Caucasian	14	11	
African American	6	7	
Latin American	2	3	
Asian American	1	2	
WRMT-R Word Identification[a]	59.52 (7.61)	62.30 (13.36)	−1.12
Grade equivalent	3.7	3.9	
WRMT-R Word Attack[a]	23.35 (6.42)	24.30 (9.83)	−0.50
Grade equivalent	2.8	2.9	
PPVT-III[b]	101.83 (12.53)	102.13 (16.81)	−0.09
WRMT-R Comprehension[b]	100.09 (7.12)	101.09 (11.44)	−0.55
Grade equivalent	3.0	3.0	
GMRT4 Comprehension[c]	47.13 (17.31)	48.48 (18.83)	−0.33
Grade equivalent	3.5	3.6	

Note. There were 23 students in each group. Values in parentheses represent the standard deviation. All values were nonsignificant at the $p < .05$ level. PPVT-III = Peabody Picture Vocabulary Test, third edition; WRMT-R = Woodcock Reading Mastery Test-Revised; GMRT4 = Gates-MacGinitie Test of Reading Comprehension, Level 3.
[a] Raw scores. [b] Standard scores; the mean for norming samples on the WRMT-R Comprehension test is 100. [c] Extended scale scores.

Table 2. Mean Performance of the Metalinguistic Treatment Group and Control Group on Pretests and Posttests, Test Statistics, and Effect Sizes

Test	Treatment	Control	*F*	*d*
Homonyms (10 maximum)				
Pretest[a]	2.78 (2.28)	2.83 (2.53)	[G] 2.01*	
Posttest	5.22 (2.65)	3.35 (2.66)	[T] 15.39***	.70
Gain	+2.44	+0.52	[G x T] 6.44**	
Ambiguous sentences (8 maximum)				
Pretest[a]	1.00 (1.35)	1.87 (1.96)	[G] 0.20*	
Posttest	2.91 (1.70)	1.61 (1.90)	[T] 22.92***	.72
Gain	+1.91	−0.26	[G x T] 39.68***	
Riddles (5 maximum)				
Pretest[a]	2.61 (1.23)	2.87 (1.49)	[G] 0.06*	
Posttest	2.74 (1.51)	2.65 (1.70)	[T] 0.03*	.06
Gain	+0.13	−0.22	[G x T] 0.47*	
Heteronyms (10 maximum)				
Pretest[a]	3.87 (2.12)	4.30 (2.82)	[G] 0.13*	
Posttest	4.13 (1.91)	4.17 (2.76)	[T] 0.07*	−.01
Gain	+0.26	−0.13	[G x T] 0.58*	
Self-corrections (%)				
Pretest[a]	14% (0.12)	19% (0.13)	[G] 0.29*	
Posttest	25% (0.18)	23% (0.15)	[T] 9.23***	.12
Gain	+11%	+4%	[G x T] 1.74*	
Story anomalies (10 maximum)				
Pretest[a]	6.13 (2.88)	6.04 (3.28)	[G] 1.02*	
Posttest	7.43 (2.95)	5.70 (3.75)	[T] 2.52*	.51
Gain	+1.30	−0.34	[G x T] 7.50***	
GMRT4 multiple-choice recognition[b]				
Pretest	468.0 (33.7)	470.8 (37.1)	[G] 0.01*	
Posttest	466.8 (31.8)	465.9 (44.3)	[T] 0.77*	.02
Gain	−1.20	−4.90	[G x T] 0.28*	
Posttest (GE)	3.4	3.4		
WRMT-R passage completion[c]				
Pretest	100.09 (7.12)	101.09 (11.44)	[G] 1.30*	
Posttest	105.22 (8.32)	97.96 (11.33)	[T] 1.42*	.73
Gain	+5.13	−3.13	[G x T] 24.23***	
Posttest (GE)	3.5	2.9		

Note. $n = 23$ per group; $df = 1, 44$. Effect sizes were calculated on posttest means: $d = M_1 - M_2/\sigma_{pooled}$, where $\sigma_{pooled} = \sqrt{[(\sigma_1^2 + \sigma_2^2)/2]}$. G = test of group (treatment vs. control); T = test of time of test (pretest vs. posttest); G x T = test of interaction between group and time of test; GMRT4 = Gates-MacGinitie Test of Reading Comprehension, Level 3; GE = grade equivalent (scores based on nationally normed sample); WRMT-R = Woodcock Reading Mastery Test-Revised.

[a] Matched-pair *t* tests were conducted to verify that the two groups did not differ statistically on pretests. The *t* statistics ($df = 22$) were as follows (all values nonsignificant): homonyms $t = -.08$; ambiguous sentences $t = -1.97$; riddles $t = -.81$; heteronyms $t = -.63$; self-corrections $t = -1.27$; story anomalies $t = .11$.

[b] Extended scale scores.

[c] Standard scores ($M = 100$, $SD = 15$) based on nationally normed sample.

* *p = ns.* ** $p < .05$. *** $p < .01$.

in both groups. Students often requested a turn with the investigator, and none ever refused to participate. During training, students showed enthusiasm in several ways—by reporting to the investigator that they had shared the training activities at home with family members, by interacting energetically, and by expressing disappointment at the end of a session.

Effectiveness of Instruction to Teach MA

Success During Training (SDT)

Students' responses to metalinguistic training in ambiguities was analyzed to understand its impact and to identify sources of difficulty. In Session 1, students were taught to pick out homonyms from a larger word set and explain the two meanings of each word. At the outset, none knew what a homonym was or where to begin. The concept was explained and students were taught two strategies for detecting homonyms, one focusing on multiple form classes and the other on synonyms. Students' attempts to apply these strategies were scaffolded by the experimenter until they could respond independently to the words. All of the students were successful at identifying some homonyms independently (*M* = 5.13 words correct, *SD* = 2.7, out of a maximum 22). At the beginning of the next session, when asked whether they remembered what a homonym was, all were able to explain the concept. These observations indicate that the task of generating multiple meanings of isolated words was relatively unfamiliar and training was effective in teaching third graders how to do this.

In Session 2, participants were taught to explain alternative meanings of seven ambiguous sentences, each illustrated with two drawings. They then used Colorforms to depict alternative meanings of eight ambiguous sentences. Although manipulating Colorforms was more interesting than was explaining sentences, it was also more distracting, as a few students moved off-task to build their own scenes. The task proved difficult, and only six students were able to construct two meanings for any of the sentences independently at first try. They could build the first meaning but the experimenter had to scaffold students' construction of the second meaning. Scaffolding involved analyzing another sentence having the same form (e.g., "The baseball bat was found by the monkey," "The seashell was found by the penguin"). This helped students independently create the alternative scene of the target sentence. The number of sentences for which students independently built two meanings, with or without analyzing a sample sentence of the same form, was scored (*M* = 1.54, *SD* = 1.2, out of 8 maximum).

Session 3 began with the experimenter reading riddles aloud and explaining them. Then participants wrote their own riddles by working with homonyms related to topics of special interest (e.g., baseball). Most of the students were enthusiastic about making up personalized riddles and wanted to share their riddle notebooks with others. All of the children independently made up at least two riddles that made sense (*M* = 2.21, *SD* = 1.7). An example of a student-generated baseball riddle was "Where is Derek Jeter's home? On the baseball field!" (displaying this student's understanding that "home" can be a place where you live or a base to touch in baseball). The most riddles any child wrote was seven. These observations indicate that students got the idea of how to write riddles and found this task especially enjoyable.

In Session 4, participants read two Amelia Bedelia books (Parish, 1979, 1988) containing many homonyms that are misunderstood by Amelia. For example, when told Mrs. Rogers is throwing a wedding shower, Amelia gets out the hose and sprays everyone with water. As they read, the students expressed recognition of the relevance of the books to the training sessions. Many spontaneously stopped to explain Amelia's comprehension difficulties. For those who did not stop at ambiguity points, the experimenter interrupted and asked students to explain.

The second book chronicled the professions of Amelia's various family members, for example, her cousin the boxer. After the expected meaning was read but before Amelia's alternative interpretation was shown (i.e., her cousin packs boxes), students guessed what she would say. After reading the book and predicting the second meaning for eight professions, students brainstormed additional professions with multiple meanings, and they drew pictures to illustrate both meanings. The number of Amelia's interpretations correctly predicted was scored (*M* = 3.08, *SD* = 1.5, out of 8 maximum), as were the number of additional professions with dual meanings identified independently and illustrated (*M* = 2.83, *SD* = 0.70). The most any child produced of the latter was five. These observations indicate that children understood that words can have two meanings.

Scores were added across the four sessions to yield a composite SDT measure. Scores ranged from 5 to 27 (*M* = 14.8, *SD* = 6.3), indicating substantial variation among students in their success. (Calculation and summation of z scores rather than SDT raw scores made little difference, as the two measures were highly correlated, $r = .98$.)

Posttests

Posttests were given to assess whether instruction was effective in teaching students to detect ambiguities following training. Two-way ANOVAs were conducted to compare gains of the experimental and control groups from pretest to posttest. The independent variables were treatment and time of test. Analyses were applied to performance in three MA tasks. Table 2 reports mean performance, test statistics, and effect sizes.

Significant main effects of time of test as well as significant interactions between time of test and treatment were found for two of the three MA tasks: homonym definition and ambiguous-sentence detection. From Table 2, it is apparent that students who received ambiguity training improved much more from pretest to posttest than control students did. The trained students produced more multiple meanings of homonyms as well as more multiple meanings of ambiguous sentences than did the control group. Comparison of trained students' pretest and posttest scores revealed that 83% showed gains in defining homonyms, and 91% showed gains in explaining ambiguous sentences. These results reveal that metalinguistic training was effective in teaching most students to think flexibly about the meanings of words and sentences.

No significant effects were found for riddles, a recognition task requiring students to choose between two punch lines. Chance performance was 2.5 items correct. From Table 2, it is apparent that scores hovered around chance level. To test whether mean scores were greater than chance, 2.5 was subtracted from each student's riddle score on pretests and posttests. *T* tests conducted on difference scores in each group on each test revealed no significant effects (for all $p > .05$), indicating that mean performance was no higher than chance. Although many students guessed at the answers, 30% of the students in the treatment group recognized at least four correct punch lines, indicating that some children were able to identify riddles. The fact that many students performed no higher than chance may explain why the reliability of this task was lower than that of the other assessments (i.e., $\alpha = .65$).

Transfer of Metalinguistic Instruction to Facilitate CM

To determine whether metalinguistic training facilitated CM, two-way ANOVAs were applied to the three posttests assessing various forms of language monitoring. Test statistics are reported in Table 2. Only the task involving the detection of anomalies in stories, the one considered the best measure of CM, revealed an effect of training, indicated by a significant interaction between treatment and test point. From Table 2, it is apparent that trained students made greater gains from pretest to posttest in detecting discrepancies in the story during their think-alouds than did control students. Whereas mean scores of trained students increased, mean scores of control students decreased slightly from pretest to posttest. Gains were evidenced in the majority (78%) of the trained students.

The anomaly-detection task consisted of two types of items that required different forms of knowledge to detect the anomaly. Macrostructure items were those that required students to draw on their background knowledge to recognize the inconsistencies, whereas microstructure items required students to remember what they had just read to detect the discrepancy. The question of interest was whether ambiguity training might enhance sensitivity to one item type more than to the other. Students in the treatment and control conditions had been matched on reading-comprehension pretests, so matched-pair *t* tests were conducted. Results revealed no significant differences between the two groups on either item type on the posttest (for all $p > .05$). The absence of differences could have resulted from the limited numbers of items (i.e., four microstructure and six macrostructure items). The mean percentages of items correct were 65% (macrostructure) and 68% (microstructure), revealing that neither type was easier.

On the other two monitoring tasks (i.e., heteronym pronunciation and miscue self-correction), no significant main effects or interactions involving treatment were detected (see Table 2 on page 308). In the oral reading of passages, trained students were not more likely than were controls to self-correct their word-substitution errors. In the heteronym task, trained and control students did not differ in varying the pronunciation of heteronyms correctly as they read the sentences. One possible reason for the lack of differences is that these two types of monitoring skills bore minimal relation to the ambiguity-detection skills that were taught, in contrast to the anomaly-detection task. Because students' sensitivity to multiple meanings of words and sentences had been heightened by training, they may have been more attuned to detecting discrepancies in text meanings in this task. Also, the anomaly task has more face validity as a measure of CM than does the other two tasks.

Item Analysis to Verify Metalinguistic Treatment Effects

For the three posttests that showed significant effects of ambiguity training, an item analysis was conducted. The proportions of participants who got each test item correct in the treatment and control groups were compared to determine whether the positive effects detected for participants also held across items. Table 3 shows that the ambiguity-trained group outperformed the control group in supplying multiple meanings of *every* word on the homonym-definitions posttest, hence precluding the need for a statistical test to confirm a treatment effect. Also evident in Table 3, students in the treatment group were more successful than were controls in explaining all but one of the ambiguous sentences. A paired-sample *t* test confirmed a significant difference, $t(7) = 3.97$, $p < .005$. On the anomaly-detection task, greater proportions of trained students than controls detected all but 2 of the 10 inconsistencies in the story. A paired-sample *t* test revealed that the comparison was statistically significant, favoring the treatment over the control group, $t(9) = 4.27$, $p < .002$. These findings show that almost all of the test items on these tasks (80% to 100%) were sensitive to the effects of ambiguity-detection training. Comparison of mean percentages of items correct in the three tasks suggests that detecting anomalies in stories was the easiest task for students ($M = 68\%$ correct across items), whereas explaining multiple meanings of ambiguous sentences was the hardest task ($M = 29\%$ correct).

To summarize, these findings indicate that instruction in ambiguity detection was successful in improving

Table 3. Proportion of Treatment and Control-Group Participants Who Responded Correctly to Each Item on the Posttests

Test item	Treatment	Control	Difference
Homonyms			
Dance (to move one's feet or body; a successive group of rhythmical steps or bodily motions)	22%	13%	+9%
Pen (an instrument for writing with ink; an enclosure used for confinement or safekeeping)	39%	30%	+9%
Flower (the blossom of a plant; to blossom)	39%	26%	+13%
Fish (an animal that lives in water; angling for a fish)	43%	39%	+4%
Light (something that makes things visible; not heavy)	48%	35%	+13%
Bug (an insect; to annoy)	48%	35%	+13%
Bear (a large mammal; to suffer, endure)	61%	39%	+22%
Poor (lacking money; pitiable)	61%	39%	+22%
Order (an authoritative direction; methodical or harmonious arrangement)	74%	39%	+35%
Star (a heavenly body appearing as a fixed luminous point in the night sky; to feature)	87%	43%	+44%
Ambiguous sentences			
Bouncing balls can make people laugh. (the act of bouncing a ball; a ball that is bouncing)	13%	13%	0%
The woman saw the broken cups and dishes. (cups and dishes can be broken; just cups can be broken)	22%	0%	+22%
The fat soldier's wife was standing by the window. (soldier was fat; soldier's wife was fat)	22%	13%	+9%
The man held the pipe. (a plumbing pipe; a smoking pipe)	26%	13%	+13%
The children showed the man the straw. (a drinking straw; straw for horses)	35%	17%	+18%
The girl tickled the baby with the stuffed animal. (the girl used a stuffed animal to tickle; the baby had a stuffed animal)	35%	26%	+9%
The nurse looked over the chart. (the nurse read the chart; the nurse peered above it)	65%	26%	+39%
The children saw a bat lying by the fence. (a baseball bat; an animal bat)	74%	52%	+22%

students' metalinguistic ability to identify multiple meanings of homonyms and ambiguous sentences and their CM ability to detect anomalies embedded in stories that they read. These findings support our hypotheses that MA can be taught effectively and that it transfers to CM of the type that requires the detection of semantic inconsistencies in text.

Transfer of Metalinguistic Instruction to Facilitate Reading Comprehension

To determine whether instruction in ambiguity detection enhanced third graders' reading comprehension, two-way ANOVAs were conducted to compare gains of the treatment and control groups from pretest to posttest on two reading-comprehension tests. Table 2 reports mean performance, test statistics, and effect sizes.

The WRMT-R Reading Comprehension subtest assessed students' ability to read short cloze passages aloud and provide the missing word or phrase (Woodcock, 1987, 1998). In the ANOVA, neither treatment nor time of test showed a significant main effect. However, the interaction of these two independent variables was significant. As is evident in Table 2 on page 308 mean scores of students who received ambiguity-detection training improved from pretest to posttest, whereas means of control students declined slightly, indicating that the treatment was effective in boosting reading comprehension. The GMRT4 Reading Comprehension test assessed students' ability to read longer passages silently and answer multiple-choice questions (MacGinitie et al., 2000). In the ANOVA, no significant main effects or interactions were detected, indicating that the treatment did not improve performance on this measure of reading comprehension.

It was not the case that the GMRT4 assessed totally different reading-comprehension processes than did the WRMT-R, as the two tests were moderately correlated (i.e., r = .69 on pretests, and r = .60 on posttests, for all $p < .01$), indicating substantial shared variance. From these findings, we conclude that ambiguity training improved students' reading comprehension on a measure that required students to read short passages and complete their meanings orally.

Correlations of Special Interest

Correlation coefficients between pretest, training, and posttest measures are reported in Table 4. Values to the right of the diagonal reveal relationships between measures on the entire sample ($N = 46$). Values to the left reveal relationships for the MA treatment group only ($n = 23$).

A question of interest was whether training was more effective with higher or lower ability readers. Yuill (1998) reported that less-skilled comprehenders made slightly but not significantly greater gains as a result of metalinguistic training than did skilled comprehenders. To examine this possibility, we calculated correlations between pretest measures of reading comprehension and two indicators of training effectiveness: SDT scores and gains from pretest to posttest on the MA posttests (homonym and sentence ambiguity) and the story anomaly-detection posttest. Gains were examined on measures shown earlier to benefit from ambiguity training to assess improvement that resulted specifically from training. Hypotheses were tested with an α of .05 on the group of trained students ($n = 23$). Correlation coefficients are shown in Table 5.

Results revealed that reading-comprehension pretest performance predicted SDT scores (see Table 5). These findings indicate that students with higher reading skills were more successful in responding to the metalinguistic-training activities. However, when correlations were calculated between the reading-comprehension pretests and gains from pretest to posttest, the opposite finding proved significant in two of the six relationships. GMRT4 reading comprehension predicted gains in the homonym task, and WRMT-R reading comprehension predicted gains in the story anomaly-detection task (see Table 5). These correlations indicate that poorer comprehenders improved more than did better comprehenders in defining multiple meanings of ambiguous words and in detecting semantic inconsistencies in text as a result of training. The negative correlations did not result from ceiling effects suppressing gain scores of better comprehenders. These findings provide some support for Yuill's (1998) study, which suggested that ambiguity training might exert a greater impact on poorer comprehenders.

Another question of interest was whether the three MA tasks involving ambiguity detection measured the same construct and likewise whether the three CM tasks assessed the same construct. Correlations were calculated on the posttest scores of all students ($N = 46$) and tested with an Alpha of .01. Results revealed that the three MA tasks (homonyms, ambiguous sentences, riddles) were significantly related to each other, with all r values ranging from .45 to .74 (see Table 4). Although the riddles task had acceptable but lower reliability and children's mean performance was at chance level, there was sufficient variability in scores extending above chance to support positive correlations with the other MA measures.

In contrast, the three CM tasks (heteronyms, story anomalies, self-corrections) were not significantly related to each other, with rs ranging from .10 to .33 (see Table 4). However, two of the CM tasks were significantly related to all three of the MA tasks: story-anomaly detection, with rs ranging from .49 to .50, and heteronym pronunciation, with rs ranging from .47 to .53. The self-correction measure was not significantly correlated with any of the MA, CM, or reading-comprehension measures (see Table 4). These findings suggest that five of the six tasks were tapping similar processes. This suggests a close relationship between MA involving ambiguity detection and CM involving text inconsistencies and heteronym pronunciations.

According to Tunmer and Bowey's (1984) model, MA tasks that involve the processing of lexical and structural ambiguity should be related primarily to reading-comprehension skills and not to decoding skills. To examine this, partial correlations were calculated—first with reading comprehension (WRMT-R posttest) partialed out and second with decoding (either WRMT-R Word Identification or WRMT-R Word Attack pretests) partialed out—and tested for significance with an Alpha level of .01 on the full sample ($N = 46$). The MA tasks were posttests assessing homonyms and ambiguous sentences but not riddles showing chance-level performance. We reasoned that if Tunmer and Bowey's (1984) claim is true, then correlations between MA and decoding should become nonsignificant when reading comprehension is partialed out. However, correlations between MA and reading comprehension should remain strong when decoding is partialed out.

Results support this prediction: Once reading comprehension was partialed out, correlations involving word decoding and either homonyms or ambiguous sentences became nonsignificant, with rs ranging from −.02 to 0.14 ($p > .01$). However, when word-decoding skills were partialed out, correlations between reading comprehension and either homonyms or ambiguous sentences remained strong and significant, with all r values ranging from .47 to .57 ($p < .01$). These findings confirm that MA involving ambiguity detection is related primarily to higher-level reading-comprehension processes rather than to lower-level word-decoding skills, thus providing support for Tunmer and Bowey's (1984) model.

Tunmer and Bowey's (1984) model also suggests that the contribution of MA training to reading

Table 4. Correlation Coefficients Between Measures for the Full Sample and for the Metalinguistic Awareness Treatment Group and Means and Standard Deviations for the Full Sample

Assessment	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	*M*	*SD*
Pretests																					
1. Homonyms	—	.53*	.23	.50*	.16	.27	.40*	.27	.42*	.47*	-	.46*	.52*	.36*	.37*	.15	.17	.45*	.39*	2.80	2.38
2. Ambiguous sentences	.53*	—	.33	.65*	.42*	.34	.64*	.44*	.61*	.59*	-	.39*	.62*	.52*	.58*	.20	.23	.39*	.58*	1.43	1.72
3. Riddles	.33	.33	—	.55*	.21	.28	.39*	.17	.45*	.37*	-	.29	.38*	.33	.41*	.25	.32	.35	.29	2.74	1.36
4. Heteronyms	.64*	.56*	.43	—	.27	.41*	.77*	.58*	.74*	.63*	-	.42*	.57*	.49*	.74*	.35	.35	.58*	.64*	4.09	2.48
5. Self-corrections	.35	.42	.04	.22	—	.22	.28	.17	.26	.18	-	.27	.19	.18	.37*	.35	−.02	.02	.20	0.16	0.12
6. Story anomalies	.35	.40	.30	.48	.17	—	.51*	.40*	.49*	.53*	-	.50*	.49*	.39*	.42*	.21	.78*	.50*	.48*	6.09	3.05
7. WRMT-R Word Identification	.55*	.58*	.24	.65*	.10	.44	—	.82*	.80*	.72*	-	.42*	.47*	.61*	.75*	.16	.37*	.60*	.83*	60.91	10.84
8. WRMT-R Word Attack	.24	.24	−.01	.27	.13	.26	.56*	—	.65*	.51*	-	.25	.31	.42*	.64*	.15	.28	.45*	.70*	23.83	8.23
9. WRMT-R Reading Comprehension	.48	.53*	.08	.58*	.26	.64*	.74*	.46	—	.69*	-	.38*	.52*	.49*	.72*	.24	.36*	.76*	.68*	100.59	9.43
10. GMRT4 Reading Comprehension	.58*	.76*	.22	.64*	.18	.56*	.78*	.34	.68*	—	-	.43*	.53*	.52*	.48*	.15	.39*	.65*	.80*	469.41	35.07
Training																					
11. Success during training	.39	.69*	.34	.62*	−.05	.47*	.77*	.40	.53*	.72*	-	-	-	-	-	-	-	-	-	15.00[a]	6.35[a]
Posttests																					
12. Homonyms	.55*	.38	.29	.27	.31	.50	.39	.20	.43	.30	.37	—	.74*	.45*	.47*	.05	.49*	.59*	.39*	4.28	2.79
13. Ambiguous sentences	.50	.75*	.46	.51*	.27	.56*	.51*	.27	.49	.53*	.71*	.55*	—	.55*	.54*	.28	.50*	.63*	.53*	2.26	1.90
14. Riddles	.50	.71*	.36	.37	.29	.43	.58*	.09	.36	.51*	.61*	.56*	.70*	—	.49*	.17	.50*	.53	.65*	2.70	1.59
15. Heteronyms	.39	.44	.25	.60*	.22	.50	.69*	.36	.69*	.50*	.49	.56*	.48	.39	—	.33	.29	.54*	.59*	4.15	2.35
16. Self-corrections	.20	.28	.25	.40	.42	.21	−.06	−.11	.21	−.03	−.02	−.02	.36	.18	.11	—	.10	.08	.15	0.24	0.16
17. Story anomalies	.33	.25	.35	.29	−.05	.80*	.16	.01	.26	.34	.30	.35	.35	.41	.09	.09	—	.50*	.40*	6.57	3.45
18. WRMT-R Reading Comprehension	.45	.56*	.05	.53*	.05	.59*	.59*	.27	.74*	.69*	.61*	.54*	.44	.41	.57*	−.11	.40	—	.60*	101.59	10.49
19. GMRT4 Reading Comprehension	.35	.69*	.08	.52*	.16	.34	.78*	.41	.51*	.78*	.75*	.15	.53*	.59*	.45	−.06	.12	.43	—	464.63	40.24

Note. Correlation coefficients to the right of the diagonal represent the full sample ($N = 46$), and correlation coefficients to the left of the diagonal represent the metalinguistic-awareness treatment group ($n = 23$).
[a] Mean and standard deviation for the metalinguistic awareness treatment group.
* $p < .01$.

Table 5. Correlation Coefficients Between Pretest Measures and Gains From Pretest to Posttest on Three Posttests (n = 23)

Assessment	1	2	3	4	5
1. WRMT-R Reading Comprehension					
2. GMRT4 Reading Comprehension	.69**				
3. Success during training	.53**	.72**			
4. Homonym gain	–.25	–.52*	–.21		
5. Ambiguous-sentences gain	.10	–.16	.22	.21	
6. Story-anomalies gain	–.58**	–.34	–.25	.01	–.21

* $p < .05$. ** $p < .01$.

comprehension is mediated by a CM strategy. To test this possibility, partial correlations were calculated. If true, then partialing out scores on the CM story anomaly-detection posttest should eliminate the significant relationships between MA (homonym and ambiguous-sentence posttests) and reading comprehension (WRMT-R posttest). Results provide little support for this hypothesis, however. Partial correlations remained statistically significant (N = 46, for all p < .01). For the homonym posttest, r = .59 without partialing and r = .46 with partialing. For the ambiguous-sentence posttest, r = .63 without partialing and r = .51 with partialing. These findings suggest that MA made a direct contribution to reading comprehension and that it did not require CM to mediate its effects.

Discussion

To summarize, results showed that MA involving ambiguity detection could be taught effectively to third graders. Those receiving MA instruction improved in their ability to give multiple definitions of ambiguous words and to explain double meanings of ambiguous sentences compared with a control group. The skills acquired during MA instruction transferred to skills not directly taught, such as those involving CM and reading comprehension. Trained students showed superior CM in a task requiring them to detect semantic inconsistencies as they read a story aloud. Also, trained students comprehended passages better than did controls on the WRMT-R Reading Comprehension subtest but not on the GMRT4 Reading Comprehension subtest. These findings were based on experimental evidence involving treatment and control groups with random assignment, thus providing support for MA instruction as the cause of these findings.

Effects of MA Instruction on MA and the Processing of Language

The effectiveness of MA instruction was evidenced not only by trained students' superior performance on homonym and sentence-ambiguity posttests compared with control students but also by the fact that a high proportion of trained students showed gains from pretest to posttest on these measures (i.e., 83% improved from pretest to posttest in detecting homonyms, and 91% improved in detecting ambiguous sentences). In addition, the trained students were observed to be responsive, enthusiastic, and generally successful in the activities used to teach MA during instructional sessions.

How exactly did MA instruction alter students' MA and their processing of language? Our explanation is that students' thinking about the meanings of words and sentences in and out of text became less rigid, more active, and more flexible (Cartwright, 2008). Rather than terminating thought after deriving one interpretation of verbal information, students learned to remain open to alternative interpretations and to detect sources of ambiguity, as Cairns et al. (2004) suggested. Several of the training tasks required this type of processing: having to think of two very different meanings for homonyms and ambiguous sentences, figuring out riddles and then making up new riddles that hinged on homonyms, predicting unexpected second meanings of ambiguous terms in the Amelia Bedelia books, identifying and elaborating additional ambiguous terms having expected and unexpected meanings.

The homonym and ambiguous-sentence posttests showed that trained students had learned to detect ambiguities in words and sentences. The CM posttest assessing anomaly detection also showed effects of training. This task required students to detect semantic inconsistencies in familiar stories (e.g., "'I'm a big bad wolf!' said the hungry bear"). These CM processes may have benefited from what students learned when they read the Amelia Bedelia stories and anticipated farfetched interpretations of ambiguous terms. Because Amelia's alternative interpretations were inconsistent with the conventional, expected interpretations, this may have primed students to recognize other types of semantic inconsistencies in the posttest stories and hence may explain why MA training boosted CM in the anomaly task.

How is the absence of training effects on other MA and CM posttests to be explained? The one MA task showing no effects was the riddles task, which required selecting one of two punch lines to create a riddle. Other findings indicated that this posttest lacked sensitivity in assessing what most children had learned about riddles. There were only five items, and mean performance did not rise above chance level, indicating that many students were guessing. Also, rather than choose the punch line that created a riddle, some students were observed to choose the other punch line because they regarded it as funny. For example, when asked the riddle "Why did the skeleton go to the movies by himself?" and given the punch line options of "Because he was lonely" or "Because he had no body to go with him," the idea of describing a skeleton as "lonely" made many students giggle, even though the punch line that contains multiple meanings plays on the division of the word *nobody*. Although the riddles task appeared to be too insensitive to detect effects of the MA treatment, nevertheless performance was positively correlated with the other measures of MA, indicating that students who scored above chance level possessed greater awareness of semantic ambiguities. The possibility that a better-designed riddles task would show the benefits of MA training awaits future study.

The posttests of heteronym pronunciation and miscue self-correction were included to assess CM because they involved language monitoring at the word level. Performance on these tasks showed no benefit of metalinguistic instruction. The task assessing students' self-corrections of their oral miscues was not significantly correlated with any other pretest or posttest, including the other CM tasks. Self-correction of miscues may not be a good indicator of CM because miscues may be corrected or not corrected for other reasons. They may go uncorrected when they are consistent with the meaning of the text, and they may be corrected when letters in the written words do not match sounds in the substituted words. The heteronym task required students to read sentences and pronounce a repeated spelling in two different ways to reflect different meanings (e.g., "I found a *live* bear where I *live*"). Analysis of the processes underlying success in this task suggests that detection of ambiguity may not be required. Simply comprehending the intended meanings of the heteronyms in the sentences may activate correct pronunciations. In fact, readers who fail to notice that the terms are ambiguous and hence are not distracted by them might do even better. Thus, it may not be surprising that training in ambiguity detection did not contribute to performance in this task.

Effects of MA Instruction on Reading Comprehension

Results were partially supportive of the hypothesis that MA instruction would enhance reading comprehension. Students who received MA training showed a substantial gain from pretest to posttest and outperformed the control students on a standardized reading-comprehension test (WRMT-R) that was individually administered and involved reading and filling in the missing words to complete the meanings of short passages. However, trained students did not show superior performance on a group-administered reading-comprehension test (GMRT4) that involved reading longer passages silently and answering multiple-choice questions. What explains the difference in findings is not clear and may involve either the conditions of administration or the type of comprehension task. Studies by Cutting and Scarborough (2006) and Keenan, Betjemann, and Olson (2008) indicated that different tests of reading comprehension may not tap the same array of cognitive processes.

One possible explanation for our findings is that the skills taught during MA training exerted a bigger impact on the ability to complete passage meanings than they did on the recognition of correct answers from multiple choices. Cutting and Scarborough (2006) discussed how different reading-comprehension tasks require different vocabulary- and sentence-processing abilities. Most of the WRMT-R passages consisted of two sentences with a blank space in the latter half or at the end of the passage. To arrive at their answers, students had to apply background knowledge and reasoning to connect earlier and later information in the passages and then express their understanding verbally. Metalinguistic training may have improved students' ability to focus on and talk about the meanings of passages by teaching them to regard written language as an object with parts whose meanings can be processed separately, analyzed, and reanalyzed. In contrast, the GMRT4 passages were longer, and comprehension was measured by multiple-choice questions at the end of each passage. Students merely had to recognize correct answers. They did not have to manipulate meanings to express their understanding. The impact of MA training on the ability to analyze and talk about the meanings of text merits more research.

Another possibility is that differences in the way the two tests were administered diminished the sensitivity of the GMRT4 in assessing reading comprehension. The test consisted of 11 passages, each about three paragraphs long. Students completed the test silently in groups of three or four during a separate session lasting 45 minutes. They did not have to produce any answers orally but only had to choose among answers already provided. Some students, very likely the poorer readers, appeared not to try their best, possibly because the test was too hard. They appeared to mark answers without reading the passages. They complained that there were too many passages. They looked at their neighbor's answers, despite remonstrations. They exhibited signs of fatigue. The greater presence of these behaviors during the GMRT4 test than during the WRMT-R test may explain why the former provided a less-sensitive measure of reading comprehension. Very likely, correlations between the two tests remained strong because poor readers consistently received low scores either from comprehension difficulties or from lack of effort.

The finding that students trained in MA outperformed controls on a standardized reading-comprehension test is consistent with results reported by Yuill (1998). Her individually administered test of comprehension (i.e., Neale Analysis of Reading Ability [Neale, 1997]) required students to read aloud short passages and then answer questions orally from memory. Furthermore, the effect size in our study, $d = .73$ on the WRMT-R, is comparable to Yuill's effect size on the Neale reading-comprehension posttest. The effect size for students entering her study as more skilled comprehenders was $d = .53$ and those entering as less-skilled comprehenders was $d = .93$. All of these effect sizes are greater than those reported by Rosenshine and Meister (1994) in their meta-analysis of 16 studies on reciprocal teaching. On standardized reading-comprehension tests, the median effect size was $d = .32$. The fact that effect sizes resulting from MA instruction were substantially higher indicates the strength of this type of instruction.

Researchers studying how to improve reading comprehension have focused on three major causal factors: word decoding; background knowledge, including vocabulary and content knowledge; and cognitive and metacognitive strategies (Pressley, 2000). Less attention has been paid to MA. Present findings suggest that this may be an oversight. Language lies at the core of reading comprehension. To understand and interpret text, readers must engage in active thinking and verbalizing about the meanings of words and sentences, must remain flexible in discerning intended meanings, and must engage the psycholinguistic processing skills that they employ in aural comprehension. This processing is an integral part of what readers do to construct a text's meaning; it contrasts with the processing involved in the application of cognitive and metacognitive strategies during text reading, for example, creating mental images of meanings, summarizing the main idea, or asking questions about the text. The latter processing is external and optional in its application rather than central to the construction of text representations. Present findings underscore the importance of pursuing research on MA to advance our understanding about meaning construction during reading comprehension.

The reason that CM tasks were included was to determine whether effects of MA training on reading comprehension might be mediated by a CM strategy. According to Tunmer and Bowey (1984), the contribution of MA to reading comprehension is that it gives children conscious access to their implicit linguistic knowledge when they are reading and thereby enables them to better monitor their comprehension and make repairs when necessary. However, findings did not provide much support for this idea. The three tasks assessing the CM construct were not correlated, raising doubt that they measured a single construct. The fact that two of the CM tasks (heteronyms and story anomalies) were strongly correlated with the three MA tasks is consistent with the possibility that CM might mediate effects of MA on reading comprehension. However, when effects of CM were controlled using partial correlations, results revealed that MA was still strongly correlated with reading comprehension. This suggested that training in MA exerted a direct impact on reading comprehension, precluding the need to consider any mediator.

Strengths and Limitations

The design of the study allowed us to rule out alternative explanations for the effects of metalinguistic training on posttest performance. Preexisting group differences were ruled out by matching students and randomly assigning members of pairs to treatments and by administering pretests to verify that the groups did not differ prior to training. Hawthorne effects were ruled out. The control treatment was designed to give students special attention outside of the classroom and to engage and motivate them as much as the experimental group. Students worked in groups as they read and discussed a story known to interest third graders. Observations confirmed that students were interested, motivated, and convinced that they were receiving beneficial instruction. Their classroom teachers also reinforced this belief.

The design of the study used a no-treatment control group to assess whether metalinguistic instruction contributes to reading comprehension. It remains for future researchers to determine whether this type of training is more effective than are alternative types of instruction. In addition, the effects of metalinguistic instruction were assessed when the training ended. It remains unclear whether effects would persist over time and, if not, whether more extensive instruction might be needed to ensure more lasting effects. These possibilities merit further study.

The populations of participants sampled here and in Yuill's (1998) study were somewhat narrow. Yuill worked with 7- and 8-year-olds enrolled in British schools in the United Kingdom. Students in the present study were lower socioeconomic status third graders from a variety of cultural backgrounds enrolled in U.S. urban schools. Whether the metalinguistic activities taught in these studies would be effective with students in other grades from other backgrounds awaits study. However, given the reading-comprehension difficulties of lower socioeconomic status students, it is important to identify forms of reading instruction that are potentially beneficial for these students.

The tasks used to assess MA were devised by the experimenter. One of these tasks fell short in its sensitivity. Mean performance on the riddles task was no higher than chance, possibly because there were too few items (only five), with only two-choice answers. However, correlations indicated that some students who were taught about riddles during training obtained scores beyond chance, and riddle scores were correlated with the other MA tasks. This indicates that, with revision, the riddles task has the potential for becoming a better measure of MA.

The three CM tasks showed very low intercorrelations, raising doubt that they measured one type of monitoring. Analysis of the processing required in the three tasks revealed that the story anomaly-detection task had the strongest face validity as a measure of CM, and it was the only one of the three that showed an effect of MA instruction. These findings suggest that the story anomaly-detection task may be a good measure of CM, whereas the heteronym and self-correction measures tap other forms of language monitoring that are less indicative of comprehension processes.

Although we have interpreted performance on our tasks to bear on issues of construct validity, our findings are merely suggestive. A study designed for this purpose with a larger sample of participants and tests with a greater number of items is required to provide stronger evidence and to settle uncertainties raised by this study.

Implications

The results of this study carry important implications. MA in the form of phonological awareness is widely recognized to help children learn to read, but MA that involves processing multiple meanings and detecting ambiguities has yet to be recognized as a facilitator of reading. Present findings suggest that this new direction in MA research may hold promise for extending our understanding of text-comprehension processes and for improving the effectiveness of reading-comprehension instruction in the early elementary grades.

Ambiguity detection qualifies as a type of MA in that students must consciously wield control over their mental processes to recognize that words and sentences have double meanings and to reprocess those meanings. One question for future researchers is whether this form of MA facilitates general language processing in listening-comprehension as well as reading-comprehension tasks or whether its effects are limited to reading.

Ambiguity-awareness research carries possible implications for testing, diagnosis, and early intervention. Cairns et al. (2004) showed that lexical ambiguity-detection skill measured at the beginning of first grade predicted later reading ability in second grade, and both lexical and structural ambiguity-detection skill assessed in second grade predicted third-grade reading ability. Shakibai (2007) demonstrated that it is possible to teach kindergarten children to detect homonyms and lexical ambiguities. The present study indicates that teaching third graders homonym- and ambiguity-detection skills (both lexical and structural) improves their reading comprehension. Combined results of these studies suggest that homonym- and ambiguity-detection tasks might be useful in identifying beginning first graders who are at risk for reading difficulty and in designing effective reading-comprehension instruction for them. Training in homonym and ambiguity detection could be incorporated into emergent literacy programs for prereaders and into intervention programs for older struggling readers with beneficial effects. These possibilities await attention.

This research also holds implications for reading-comprehension instruction. Present findings suggest that teaching children to manipulate language, write riddles, and read ambiguous text, such as the popular children's series Amelia Bedelia, increases their understanding of ambiguity and their reading comprehension. The methods used here could easily be adapted to whole-class instruction. The enthusiasm exhibited by students in the present study indicates that these methods would be popular and enjoyable. In sum, present findings suggest that a significant way to expand the design of reading-comprehension instruction may be by teaching children how to recognize and think about the ambiguities in language.

Notes

This study was conducted as a PhD dissertation by the first author. We thank David Rindskopf for his helpful comments and suggestions. We are grateful to students and staff at the public elementary schools in New York City, where this study was conducted.

References

Baker, L., & Zimlin, L. (1989). Instructional effects on children's use of two levels of standards for evaluating their comprehension. *Journal of Educational Psychology*, *81*(3), 340–346. doi:10.1037/0022-0663.81.3.340

Bernstein, J.E. (1979). *Fiddle with a riddle*. New York: E.P. Dutton.
Cairns, H.S. (1999). *Psycholinguistics: An introduction*. Austin, TX: Pro-Ed.
Cairns, H.S., Waltzman, D., & Schlisselberg, G. (2004). Detecting the ambiguity of sentences: Relationship to early reading skill. *Communication Disorders Quarterly*, *25*(2), 68–78. doi:10.1177/15257401040250020401
Carlisle, J. (1995). Morphological awareness and early reading achievement. In L. Feldman (Ed.), *Morphological aspects of language processing* (pp. 189–210). Hillsdale, NJ: Erlbaum.
Cartwright, K.B. (Ed.). (2008). *Literacy processes: Cognitive flexibility in learning and teaching*. New York: Guilford.
Cutting, L.E., & Scarborough, H.S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, *10*(3), 277–299. doi:10.1207/s1532799xssr1003_5
Dunn, L.M., & Dunn, L.M. (1997). *Peabody Picture Vocabulary* (3rd ed.). Circle Pines, MN: American Guidance Service.
Ehri, L.C., Nunes, S.R., Stahl, S.A., & Willows, D.M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, *71*(3), 393–447. doi:10.3102/00346543071003393
Fowles, B., & Glanz, M.E. (1977). Competence and talent in verbal riddle comprehension. *Journal of Child Language*, *4*(3), 433–452. doi:10.1017/S0305000900001781
Gaux, C., & Gombert, J.E. (1999). Implicit and explicit syntactic knowledge and reading in pre-adolescents. *British Journal of Developmental Psychology*, *17*(2), 169–188. doi:10.1348/026151099165212
Hacker, D.J. (1997). Comprehension monitoring of written discourse across early-to-middle adolescence. *Reading and Writing: An Interdisciplinary Journal*, *9*(3), 207–240. doi:10.1023/A:1007989901667
Hakes, D.T. (1980). *The development of metalinguistic abilities in children*. New York: Springer-Verlag.
Hirsh-Pasek, K., Gleitman, L.R., & Gleitman, H. (1978). What did the brain say to the mind? A study of the detection and report of ambiguity by young children. In A. Sinclair, R.J. Jarvella, & W.J.M. Levelt (Eds.), *The child's conception of language* (pp. 97–132). New York: Springer-Verlag.
Keenan, J.M., Betjemann, R.S., & Olson, R.K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*(3), 281–300. doi:10.1080/10888430802132279
Love, T.E., Swinney, D.A., Bagdasaryan, S., & Prather, P. (1999, March). Real-time processing of lexical ambiguities by pre-school children. Paper presented at the annual CUNY Conference on Human Sentence Processing, New York, NY.
MacGinitie, W.H., MacGinitie, R.K., Maria, K., & Dreyer, L. (2000). *Gates-MacGinitie reading tests* (4th ed.). Chicago: Riverside.
Mahony, D.L., & Mann, V.A. (1992). Using children's humor to clarify the relationship between linguistic awareness and early reading ability. *Cognition*, *45*(2), 163–186. doi:10.1016/0010-0277(92)90028-G
Mahony, D.L., & Mann, V.A. (1998). Children's linguistic humor and reading research: A match made in heaven. *Cahiers de Psychologie Cognitive*, *17*(2), 287–311.
Nation, K., & Snowling, M.J. (2000). Factors influencing syntactic awareness skills in normal readers and poor comprehenders. *Applied Psycholinguistics*, *21*(2), 229–241. doi:10.1017/S0142716400002046
Neale, M.D. (1997). *Neale Analysis of Reading Ability–Revised*. London: NFER-Nelson.
Pressley, M. (2000). What should comprehension instruction be the instruction of? In M.L. Kamil, P.B. Mosenthal, P.D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 545–562). Mahwah, NJ: Erlbaum.
→ Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research*, *64*(4), 479–530.
Roth, F.P., Speece, D.L., Cooper, D.H., & De La Paz, S. (1996). Unresolved mysteries: How do metalinguistic and narrative skills connect with early reading? *The Journal of Special Education*, *30*(3), 257–277. doi:10.1177/002246699603000303
Shakibai, M.T. (2007). *The efficacy of a training program to teach kindergarteners lexical ambiguity detection*. Unpublished doctoral dissertation, Graduate Center of the City University of New York, New York.
→ Shultz, T.R. (1974). Development of the appreciation of riddles. *Child Development*, *45*(1), 100–105. doi:10.2307/1127755
Swinney, D.A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 645–659. doi:10.1016/S0022-5371(79)90355-4
Swinney, D., & Prather, P. (1989). On the comprehension of lexical ambiguity in young children: Investigations into the development of mental modularity. In D. Gorfein (Ed.), *Resolving semantic ambiguity* (pp. 225–238). New York: Springer-Verlag.
Tunmer, W.E., & Bowey, J.A. (1984). Metalinguistic awareness and reading acquisition. In W.E. Tunmer, C. Pratt, & M.L. Herriman (Eds.), *Metalinguistic awareness in children* (pp. 144–168). New York: Springer-Verlag.
Tunmer, W.E., Nesdale, A.R., & Wright, A.D. (1987). Syntactic awareness and reading acquisition. *British Journal of Developmental Psychology*, *5*(1), 25–34.
van Kleeck, A., & Reddick, C. (1982, October). *The relationship between concrete operational thought and the development of metalinguistic skill*. Paper presented at the seventh annual Boston University conference on language development, Boston, MA.
Woodcock, R.W. (1987). *Woodcock Reading Mastery Tests-Revised (WRMT-R)*. Circle Pines, MN: American Guidance Service.
Woodcock, R.W. (1998). *Woodcock Reading Mastery Tests-Revised normative update*. Circle Pines, MN: American Guidance Service.
Yalisove, D. (1978). The effect of riddle structure on children's comprehension of riddles. *Developmental Psychology*, *14*(2), 173–180. doi:10.1037/0012-1649.14.2.173
Yuill, N. (1996). A funny thing happened on the way to the classroom: Jokes, riddles, and metalinguistic awareness in understanding and improving poor comprehension in children. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 193–220). Mahwah, NJ: Erlbaum.
Yuill, N. (1998). Reading and riddling: The role of riddle appreciation in understanding and improving poor text comprehension in children. *Cahiers de Psychologie Cognitive*, *17*(2), 313–342.
Zipke, M. (2007). The role of metalinguistic awareness in the reading comprehension of sixth and seventh graders. *Reading Psychology*, *28*(4), 375–396. doi:10.1080/02702710701260615
Zipke, M. (2008). Teaching metalinguistic awareness and reading comprehension with riddles. *The Reading Teacher*, *62*(2), 128–137. doi:10.1598/RT.62.2.4

Children's Literature Cited

Lobel, A. (1977). *Mouse soup*. New York: HarperCollins.
Parish, P. (1979). *Amelia Bedelia and the surprise shower*. New York: HarperCollins.
Parish, P. (1988). *Amelia Bedelia's family album*. New York: Greenwillow.

Submitted December 10, 2007
Final revision received December 19, 2008
Accepted December 30, 2008

Marcy Zipke is an assistant professor in the Department of Elementary/Special Education, Providence College, RI, USA; e-mail mzipke@providence.edu.

Linnea C. Ehri is a Distinguished Professor in the Program in Educational Psychology, Graduate Center of the City University of New York, NY, USA; e-mail LEhri@gc.cuny.edu.

Helen Smith Cairns is a Professor Emerita of the Program in Linguistics, Graduate Center of the City University of New York, NY, USA; e-mail hcairns@optonline.net.

Appendix

Homonym Definition:

Pretest Homonyms:

1. Bow
2. Bank
3. Feet
4. Mouth
5. Sink
6. Treat
7. Check
8. Can
9. Shed
10. Run

Posttest Homonyms:

1. Star
2. Poor
3. Light
4. Bear
5. Pen
6. Flower
7. Fish
8. Bug
9. Dance
10. Order

Ambiguous-Sentence Detection:

Examples: "The chicken was ready to eat."
"They talked about the problem with the teacher."

Pretest Sentences:

1. The man's nails were very sharp.
2. The elephant was ready to lift.
3. The cold made Betty feel terrible.
4. The boy watched the little fish and turtles.
5. The glasses fell on the floor and broke.
6. The sheriff caught the man with the gun.
7. The boy picked up the bow.
8. Flying kites can be exciting.

Posttest Sentences:

1. The woman saw the broken cups and dishes.
2. The children saw a bat lying by the fence.
3. Bouncing balls can make people laugh.
4. The nurse looked over the chart.
5. The fat soldier's wife was standing by the window.
6. The children showed the man the straw.
7. The girl tickled the baby with the stuffed animal.
8. The man held the pipe.

Riddle Resolution

Pretest Riddles (correct answers labeled "a"):

1. Why should you never swim on a full stomach?
   a. It's easier to swim in water.
   b. You'll get sick.
2. Why didn't anyone take the bus to school?
   a. It wouldn't fit through the door.
   b. It wasn't cool.
3. What kind of animal can jump higher than the Empire State Building?
   a. Any animal—the Empire State Building can't jump!
   b. None.
4. How many sheep does it take to make a sweater?
   a. I didn't even know they could knit!
   b. Ten.
5. Why did the skeleton go to the movies by himself?
   a. He had no body to go with him.
   b. He was lonely.

Posttest Riddles (correct answers labeled "a"):

1. Do you sleep on your stomach?
   a. No, on a bed.
   b. Yes.
2. Why did Frog eat a lamp?
   a. He wanted to eat a light snack.
   b. He was hungry.

3. Why was the policeman in bed?
   a. He was an undercover cop.
   b. He was tired.
4. Why is a school yard larger at recess than at any other time?
   a. At recess there are more feet in it.
   b. It isn't.
5. What kind of house weighs the least?
   a. A lighthouse.
   b. An apartment.