DOCUMENT RESUME

ED 136 212 CS 003 262

AUTHOR Wagner, Michael; Rohwer, William D., Jr.
TITLE Developmental Changes in Prose Memory.
PUB DATE 77
GRANT HDO-3869
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, New York, April 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Developmental Stages; Elementary Secondary Education; Grade 5; Grade 11; *Learning Processes; *Memory; *Prose; *Psychological Studies; Reading Research; *Recall (Psychological); Retention

ABSTRACT

A recognition paradigm was employed to assess developmental changes in memory for within-sentence (premise) and between-sentence (inference) information from 16 brief prose stories. Eleventh graders retained significant amounts of both premise and inference information, while fifth graders showed substantial retention only for premise information. The discrepancies between these and previous findings are explicated with regard to instructional conditions, story construction, and test list design. It is concluded that fifth graders are less inclined to retain between-sentence information from prose materials than are eleventh graders. (Author)

Developmental Changes in Prose Memory[1]

Michael Wagner

and

William D. Rohwer, Jr.
University of California, Berkeley

Abstract

A recognition paradigm was employed to assess developmental changes in memory for within-sentence (premise) and between-sentence (inference) information from sixteen brief prose stories. Eleventh graders retained significant amounts of both premise and inference information while fifth graders showed substantial retention only for premise information. The discrepancies between these and previous findings are explicated with regard to instructional conditions, story construction, and test list design. It is concluded that fifth graders are less inclined towards retaining between-sentence information from prose materials than are eleventh graders.

[1]This research was supported in part by a grant from the National Institutes of Health (HD03869). The report was prepared at the Institute of Human Learning, which is supported by grants from the National Institutes of Health.

We are grateful to Robert Weber, to the principals, Keith Burnham, Helen McDonald, and Cynthia Scortt, to the teachers and students of the Orinda Union School District, and to the students, teachers and principal, Robert Callan, of Miramonte High School, for their participation and assistance in conducting this research.

In recent years, the notion that adults 'chunk' discourse into units of meaning or 'essential ideas' (Brent, 1969) has attracted many adherents. Consistent with this view are the ideas that, at storage, adults typically use the information in prose materials to construct semantic descriptions of situations and that this integrative process applies across sentence boundaries, resulting in psychological representations of between-sentence or inference information, at the expense of syntactic and semantic information about the boundaries of individual sentences (Bransford and Franks, 1971; Bransford, Barclay, and Franks, 1972; Barclay, 1973; Singer, 1973; Singer and Rosenberg, 1973; Barclay and Reid, 1974; Bransford and Franks, 1974). These findings have even been extended to elementary school children (Paris and Carter, 1973; Paris and Mahoney, 1974).

Despite the weight of this evidence, the aim of the present study was to verify these notions once again. The hypothesis in question is that adults are able to identify assertions as being semantically congruent or incongruent with respect to presented materials, whether the assertions contain within- (premise) or between-sentence (inference) information. The reason for retracing the work of previous investigators is that there are several methodological features common to most of the studies cited above which fail to rule out alternative explanations. Thus, one goal of this study is to eliminate such alternatives. Furthermore, if the hypothesis regarding adults is verified, then the question remains as to whether it

must be qualified to accommodate age differences in performance.

The core assertions are manifest in a study by Bransford, Barclay, and Franks (1972) who presented adults with a set of three-sentence stories, describing spatial relationships among several objects, from which inferences could be drawn. The participants were presented with sets of four assertions at test (valid and invalid premises and inferences) and were asked to choose the one they had actually heard before. In one condition the test sets retained the subject and object nouns in their original order (verbatim condition) while in the other condition the order of the subject and object nouns was reversed (paraphrased condition), making it necessary also to alter the relational term. The participants chose valid inferences as often as valid premises. Thus the results indicated that adults did not respond as if they had stored representations of individual sentences in memory but rather as if they had acquired wholistic descriptions which extended beyond the information directly expressed.

There have been several attempts to extend this integration notion to children. For example, Paris and Carter (1973) presented second and fifth graders with stories and test lists similar to those used in the Bransford et al (1972) verbatim condition. The students made old-new judgments on each of the four assertions in each test set. However, for both of the semantically valid assertions in each set no novel words were introduced, while the relational terms in the two semantically invalid assertions were novel. Since the students accepted valid inferences as often as they accepted identical assertions, it was concluded that they spontaneously integrated the semantic information in the input sentences during storage by constructing between-sentence relations.

Although Paris and Mahoney (1974) replicated these findings with second and fourth graders when using verbatim test assertions, replication failed when paraphrased test assertions were used. Both the second and the fourth graders were unable to discriminate valid from invalid assertions, whether premises or inferences, when all assertions were paraphrased; that is, when the position of the subject and object nouns was reversed relative to the input. Clearly, if the notion that children store wholistic representations of semantically related sentences is correct, performance in the original verbatim and reversed paraphrase conditions should have been equivalent. Paris and Mahoney suggest that these incongruent results may be due to the fact that the test assertions were in 'marked' and 'unmarked' linguistic forms, respectively, for the original and reversed conditions. The results, then, follow from the assumption that the 'marked' forms are more difficult than the 'unmarked' forms (Olson, 1975).

A simpler interpretation is that children are less inclined towards the retention of integrated semantic information than adults and, thus, are more inclined towards the retention of lexical, syntactic, and within-sentence semantic information. Consequently, if the children based their recognition test decisions on whether the relational term in the test assertion was old or new, the prediction for originally ordered test assertions would be congruent with the results. Furthermore, these children might have been expected to encounter considerable difficulty with the test sentences, all of which contain a number of lexical and syntactic alterations.

This methodological feature of the studies cited is not the only one that deserves closer attention. For example, in each study, the subjects were instructed to classify test assertions on the basis of whether an

identical sentence had been presented at study. However, since the intent was to assess semantic integration, rather than the ability to detect surface feature alterations in a sentence, it would seem more appropriate to encourage subjects to employ a meaning criterion. This objective might be accomplished by instructing subjects to accept assertions when they mean the same thing, or agree with, the input. Furthermore, it is possible to preclude subjects from basing their decisions on formal information alone by presenting all test assertions in paraphrased form.

Another methodological feature of the previous studies is that sets of four test assertions referring to the same input story were presented to the subjects contiguously. Consequently, each decision made by a subject could be influenced by prior related decisions. An alternative procedure is to compose a test list using only one assertion from each input story.

Finally, in the previous studies the only relational terms used were spatial in nature. As already mentioned, Paris and Mahoney (1974) offered an explanation for their subjects' poor performance on paraphrased test assertions based on the presence of 'marked' relations in these items. This interpretation suffers from the inherent difficulty of identifying 'markedness' in spatial relations across varying contexts. However, for comparative relationships (larger-smaller, stronger-weaker, etc.), the distinction is easier to draw. Therefore, half of the stories in the present study were composed of comparative relations and half of spatial relations.

The purpose of the present study, then, was to assess developmentally the semantic integration hypothesis. In particular, eleventh graders (adults) were expected to be superior to children at discriminating valid from invalid

inferences, a difference that might be explained in terms of ability, or in terms of age differences in the inclination to integrate semantic information. The experiment was designed to provide a choice between these interpretations by including instructions intended to vary inclination, that is, children (fifth graders) and adults (eleventh graders) were compared across three types of instructions (prompt, meaning, verbatim) in terms of their indentification of valid and invalid premises and inferences.

Method

Subjects

Participants in the study were drawn from a high school and two elementary schools serving the same upper-middle class suburban residential area. The older sample (Mean age = 16.67 years) was randomly selected from those eleventh graders enrolled in the required American history course who volunteered for the study. The younger sample (Mean age = 10.79 years) was chosen from among the fifth grade students attending either of two elementary schools. Seventy-two students from each grade level participated, resulting in a total sample of 144 students.

Materials

Sixteen unrelated short 'stories,' each comprised of three sentences, were constructed. The first sentence of each story, which served as a title, simply listed the three common nouns contained in the story (A, B, and C). The second sentence stated a relationship between the first two nouns mentioned in the title sentence (A relation$_1$ B) while the final sentence related the second noun with the third (B relation$_2$ C), such that the assertion, 'A relation$_2$ C,' was semantically congruent with the

story as presented. Thus, 'A relation$_2$ C' was always a valid inference assertion. The stories were constructed with the intent that both valid and invalid test assertions would be equally and highly plausible so that, in order to classify the assertions correctly, retention of the meaning of the study materials would be necessary.

The relations in eight of the stories were spatial, while those in the remaining stories were comparatives. For each story, eight types of paraphrased recognition test assertions were derived. Examples of a spatial and a comparative story and a set of the eight types of test assertions derived from a comparative story are displayed in Table 1.

Since the premise test assertions could be derived from either the first or second sentence in each story, each sentence provided two of the four premise test assertions referring to a given story. The relationships in all of the test assertions were paraphrased so as to preclude students from responding solely on the basis of whether the assertions contained a new word. In addition, to insure that students could not perform at a level above chance by simply responding to relational terms with the same meaning as the ones which appeared in the story, reversed order test assertions were included. Thus, half of both the valid and invalid test assertions contained relational terms which were synonymous with those in the input, while the other half were antonyms. In order to prevent the students from discovering that 'relation$_2$' was always the appropriate one for the inference, half of the stories were presented in BC-AB order. The

order in which a particular story was presented was counterbalanced across conditions. However, all possible test assertion types could not be counterbalanced with the order of presentation of the sentences within each story (AB-BC and BC-AB), so a systematic confounding was introduced into the study. All stories presented in the AB-BC order for half of the test lists were tested with an originally ordered assertion while the remaining stories, presented in BC-AB order, were tested with reversed order assertions. The situation was exactly reversed for the second set of four test lists.

Each recognition test list contained sixteen assertions, with each one referring to a different input story. As a result, unlike the experiments by Paris and his colleagues (Paris and Carter, 1973; Paris and Mahoney, 1974) and by Bransford et al (1972), the recognition test lists in this study contain no dependent information. All possible types of test assertions were represented in a test list for both spatial and comparative stories, with the only constraint that each half of the list was counterbalanced for story type, information, order, and validity. A single test list contained just one of the eight possible test assertions for each story so eight lists were constructed which included all possible test assertions for each story.

Order of presentation of the stories was random but was the same for all participants. The order in which the stories were tested was also randomized and the same for all participants, except that the type of test assertion chosen to test each story was different for each of the eight test lists.

Procedure

The students were randomly assigned to experimental conditions and were tested individually in a mobile laboratory. The task consisted of an acquisition and a recognition test phase. During acquisition, all students were told to listen carefully to the sixteen stories and that, afterwards, they would be tested on what they remembered. Additional instructions varied across conditions. The Prompt group was instructed to try to understand how all three objects described in each story were related to each other and were asked to verbalize these relations aloud for a spatial and a comparative practice story. After this was done for each story, the experimenter repeated the relations for the students (i.e., both within- and between-sentence relations). The instructions for the Meaning group directed the students to remember and verbalize aloud what the practice stories meant, and, after students did so, the experimenter paraphrased each story. Similarly the Verbatim group was directed to remember exactly what each story said and to repeat in a verbatim fashion the last two sentences of each story, followed by a verbatim repetition of the last two sentences of the story on the part of the experimenter. Students in all three groups were asked to continue to perform the operations they had practiced for all of the stories in the study set, but to do so covertly. Thus, all participants in the study, regardless of instructional condition, were asked to produce overt verbalizations following each practice story and covert verbalizations following each acquisition story.

Following acquisition instructions the sixteen stories were presented. They had been prerecorded on cassette tape in a male voice at the rate of

two words per second followed by a blank interval such that the total presentation time for each story was 25 seconds. The stories, which were numbered, were played on a portable tape recorder.

Following acquisition, participants in all conditions received the same test instructions. They were told to answer 'yes' if the recognition test assertion had the same meaning or agreed with what they heard in the story to which the test assertion referred. Four practice test assertions (reversed true premise, original false premise, original true inference, and reversed false inference) which referred to the two practice stories were then presented, and the responses of the students were corrected. Administration of test instructions took about two minutes. The recognition test assertions were presented at ten second intervals, and students recorded their responses, one to a page, in a test booklet.

Design

The design of the study can be summarized by designating the between- and within-subjects factors separately. In addition to the factors that were used for counterbalancing purposes, Combinations and Test lists, there were two between-subjects factors. The first, Grade (five, eleven), was included to permit an assessment of age effects, and the second, Instructions (prompt, meaning, verbatim), was included for the purpose of inclining students towards the retention of various types of information. Furthermore, the retention of different kinds of test assertions, defined by Story (spatial, comparative), Information (premise, inference), and Order (original, reversed), were evaluated as within-subjects factors and were treated as repeated measures in an analysis of variance.

Results and Discussion

Grade Effects

The variables used as an overall index of performance were the proportion of valid and invalid assertions correctly classified by the students as being true or false with respect to the study set. The results are displayed in terms of these variables in Table 2. Analysis of variance revealed a significant main effect for Grade, arising from the greater number of correct verifications made by the older students, on both valid ($F(1,96) = 12.48$, $p<.05$) and invalid ($F(1,96) = 5.09$, $p<.05$) assertions.

A more detailed appraisal of these Grade effects can be made with reference to the within-subjects factors. The first factor of interest was Information (premise and inference). There was a significant interaction of Grade x Information ($F(1,96) = 10.42$, $p<.05$) on valid assertions as shown in Figure 1. Descriptively, the form of the interaction was such that no significant Grade differences existed for premises ($F<1$), while older students correctly identified valid inferences ($F(1,96) = 23.09$, $p<.05$) more often than younger students. In fact, the younger students failed to identify valid inference test assertions more than 50 per cent of the time--the expected chance level ($F<1$).

There was also a significant interaction of Grade x Information on invalid assertions ($F(1,96) = 3.94$, $p<.05$). But here, the locus of the Grade effect was not inference ($F<1$) but premise performance: the older students correctly identified invalid premises at a significantly higher rate than the younger students ($F(1,96) = 8.83$, $p<.05$).

A limitation of the present variables as indices of performance

is that they fail to correct response bias. In fact, when a bias-free measure (hits minus false alarms) was used in the analysis, the age x information interaction failed to reach significance ($F<1$). However, this corrected measure obscures the qualitatively different patterns of results for valid and invalid assertions that produced the age x information interactions. Furthermore, as Table 2 indicates, the interactions cannot be accounted for solely in terms of grade differences in yes rates. Thus, despite the apparent appeal of a corrected index, there is reason to prefer the original variables instead.

This preferred analysis clearly supports the Bransford et al (1972) finding that adults are inclined to integrate semantic information across sentences. On a recognition test, adults can recognize assertions containing premise or inference information about equally often, so long as both types are semantically congruent with study materials. However, adults are able to identify invalid premise assertions more accurately than invalid inference assertions (See Figure 1). This would seem to provide some support for the notion, recently advanced by Lawson (1977) and Flagg (1976), that, while adults integrate semantic information across sentences, they also retain within-sentence semantic information. In contrast, the adults in the Bransford et al (1972) study found it easier to identify invalid inferences than invalid premises. This discrepancy may be due to the use of different criteria by subjects in each of the two studies. In the Bransford et al (1972) study, subjects were instructed to reject test assertions containing a meaning or a formal alteration, while in the present study a semantic alteration served as the only basis for rejection.

As for children's retention of prose, the major finding was their inability to identify valid inferences at a level significantly above chance expectations, while identifying valid premises nearly as well as eleventh graders. These outcomes suggest that children are less inclined than adults towards the integration of semantic information across sentence boundaries in prose materials. The suggestion is given added weight by the fact that the fifth-grade children correctly verified invalid premise and inference assertions at identical rates, an outcome that would be expected if the children failed to make inferences, since such failure would not diminish their ability to reject invalid inference assertions. The postulated absence of inference information in children, as compared with adults, should have been reflected in a lower overall yes rate for children than for adults on inference assertions. And, indeed, this rate for fifth graders was .43, while for eleventh graders it was .54, a significant difference ($F(1,96) = 12.14$, $p<.05$).

A remaining issue concerns the locus of the effects just described. It is possible that children actually store as much inference information as adults but forget it more rapidly. One way of examining this possibility is to reduce the memory requirements of the task by giving the participants access to the appropriate input material while they are choosing their response to each test assertion. If the locus of the previously obtained differences in performance between children and adults is due to differential forgetting, such an immediate test should reveal a different pattern of results than the delayed test. Just this kind of immediate test was administered to 48 fifth and 24 eleventh graders from

the same population that received the delayed test. The results are shown in Table 4. As can be seen, the same pattern is present in these data as was evident in the original data. Specifically, there is a low hit rate for fifth graders on valid inferences and a low false alarm rate for eleventh graders on invalid premises. Thus, it appears that the observed differences across age in the ability to verify assertions containing between-sentence information cannot be attributed to differential forgetting.

The question remains as to whether the locus of the integration effect is at storage or retrieval. Several authors (Bransford and Franks, 1971; Barclay, 1973) have claimed that the locus is at storage. Nevertheless, subjects could remember within-sentence information and use it at test to construct inferences, enabling them to correctly verify inference assertions. If the locus of integration is at retrieval, several predictions might be made. For instance, if a subject fails to remember all of the premises in the input, he should only be able to verify correctly those inferences from input stories for which he remembers both premise assertions. On the other hand, a subject should be able to identify half of the premise assertions referring to input stories, even if he remembers only a single assertion. Therefore, if integration is a retrieval phenomenon and a subject's retention of premises is less than perfect, the proportion of correct verifications of valid premise assertions should exceed that for valid inference assertions. However, in this study, eleventh graders, who presumably integrate semantic information, verified more valid inferences than premises. Although even this evidence is inconclusive, it seems to lend support to a storage interpretation of integration.

A second within-subjects factor was Order. Originally ordered test assertions were those in which the subject and object nouns remained in the same position, relative to the relational term, as when the assertion was presented at study. For valid assertions in such instances the relational term at test would simply be a paraphrase of the relation in the corresponding study sentences. In reversed test assertions, the positions of the two nouns were exchanged, so that the relational term in valid assertions was an antonym of the original. There was a significant interaction of Grade x Order ($F(1,96) = 6.82$, $p<.05$) for invalid assertions. Whereas fifth graders were able to classify reversed invalid assertions as accurately as eleventh graders, their ability to classify original invalid assertions was quite inferior (See Table 3). One possible interpretation of this effect is that, when the younger students are unsure, they tend to base their decision more heavily on whether the subject and object nouns in the assertions are in the same position relative to study than on whether the relation in the assertion has the same meaning as the one at study. Consequently, given 'A relation$_1$ B' at study, the younger students would find it easier to reject 'B relation$_1$ A' than 'A relation$_x$ B' because of the greater salience of the temporal order of occurrence of the nouns for them than of the meaning of the relational term.

The final within-subjects factor was Story, that is, whether the relations in the stories were spatial or comparative in nature. None of the interactions of Story x Grade or Story x Instructions were significant.

Instructional Effects

The effects of instructions were assessed using pairwise contrasts between the Meaning condition and each of the other two Instructional

conditions. None of these contrasts, or those associated with Grade x Instructions were significant for either valid or invalid assertions, indicating that Instructions had no effect on overall performance and no effects specific to older or younger students. The most interesting questions about Instructions, however, are concerned with effects which should have emerged in connection with the within-subjects factors.

But even here, several interactions of a priori importance, such as Instructions x Information, and Grade x Instructions x Information, failed to reach significance for either valid or invalid assertions. Thus, instructions did not appear to exert a general influence on the retention of premise or inference information for younger or older students.

There was, however, a significant interaction of Story x Grade x Instructions (meaning vs. verbatim) for valid assertions ($F(1,96) = 4.98$, $p<.05$). This was due primarily to the low proportion of valid spatial assertions correctly identified by fifth graders given Verbatim instructions. Furthermore, the interaction of Grade x Instructions (meaning vs. verbatim) x Story x Information was significant ($F(1,96) = 6.92$, $p<.05$) for valid assertions. This outcome indicates that the poor performance of fifth grade students given Verbatim instructions was for the most part limited to valid spatial inference assertions. It is interesting that the Verbatim instructions had no effects on the retention of comparative stories while at the same time depressing performance on valid inference assertions from spatial stories. Perhaps the learning of comparative stories is invariant across processing strategies as a result of some overriding influence inherent in relations of this type.

Other Effects

A separate analysis of variance was conducted to assess the effects of the factors used for counterbalancing purposes, Combinations and Test lists. Since these hypotheses were of little theoretical significance, they were tested at $\alpha = .01$. None of the main effects were significant. However, when these same factors were included in an omnibus analysis, there were a number of significant interactions which seem to reflect large variations across individual stories.

These variations may have been partly a function of the particular contextual elements of the stories. Alternatively, a possible interpretation of them, at least for the comparative stories, may be found in Clark's (1969) characterization of most comparative relations as being either 'marked' or 'unmarked.' Using Clark's definition, half of the comparative stories in this study contained 'unmarked' and the remaining half contained 'marked' terms as the primary ($relation_1$) relations. The results for this breakdown appear in Table 5. It is clear by inspection that Clark's characterization may account for a large portion of the variance between comparative stories; 'markedness' at input was a significant predictor of performance for both adults and children.

According to Clark, such differences should be primarily a result of forgetting and not variations in comprehensibility among the input stories. Consistent with this view, an examination of the immediate test data showed equivalent performance on 'marked' and 'unmarked' stories for both children and adults (See Table 5). A comparison of the immediate and delayed data revealed that for 'unmarked' input stories there was a negligible difference in performance (.01), while for 'marked' stories

there was a large difference (.20). Consistent with semantic distinction theory, this interaction was evident for both fifth and eleventh graders, and for inferences as well as premises.

To summarize, the study does not lead to the conclusion that children are less active processors of prose materials than are adults, but rather that children have available fewer or less efficient strategies which they employ in prose learning tasks. The fifth-grade students produced a high rate of correct verifications of all item types except valid inferences, suggesting that they generate many fewer inferences than adults. Even the Prompt instructions failed to facilitate children's inference performance, although some other, as yet untested, treatment might be successful in doing so.

Finally, the 'markedness' results suggest some interesting hypotheses about relationships between the structure of materials and their memorability in prose learning research. For example, equally comprehensible materials might result in either insignificant or substantial forgetting over time, depending on the degree of 'markedness' of their constituent relational terms.

References

Barclay, J. R. The role of comprehension in remembering sentences. Cognitive Psychology, 1973, 4, 229-254.

Barclay, J. R. & Reid, M. Characteristics of memory representations of sentence sets describing linear arrays. Journal of Verbal Learning and Verbal Behavior, 1974, 13, 133-137.

Bransford, J. D., Barclay, J. R., & Franks, J. J. Sentence memory: A constructive vs. interpretive approach. Cognitive Psychology, 1972, 3, 193-209.

Bransford, J. D. & Franks, J. J. The abstraction of linguistic ideas. Cognitive Psychology, 1971, 2, 331-350.

Brent, S. B. Linguistic unity, list length, and rate of presentation in serial anticipation learning. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 70-79.

Clark, H. H. Linguistic processes in deductive reasoning. Psychological Review, 1969, 76, 387-404.

Flagg, P. W. Semantic integration in sentence memory? Journal of Verbal Learning and Verbal Behavior, 1976, 15, 491-504.

Lawson, R. Representation of individual sentences and holistic ideas. Journal of Experimental Psychology: Human Learning and Memory, 1977, 3, 1-9.

Olson, D. R. On the relations between spatial and linguistic processes. In J. Eliot & N. Salkind (Eds.), Children's spatial development. Springfield, Ill.: C. C. Thomas, 1975.

Paris, S. G. & Carter, A. Y. Semantic and constructive aspects of sentence memory in children. Developmental Psychology, 1973, 9, 109-113.

Paris, S. G. & Mahoney, G. J. Cognitive integration in children's memory for sentences and pictures. Child Development, 1974, 45, 633-642.

Singer, M. The abstraction of linguistic ideas: A replication of Bransford and Franks (1971). Bulletin of the Psychonomic Society, 1973, 1, 416-418.

Singer, M. & Rosenberg, S. Role of grammatical relations in the abstraction of linguistic ideas. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 273-284.

Table 1: Examples of a Spatial and a Comparative Story (Upper Panel) and the Set of Test Assertions for the Comparative Story (Lower Panel)

	The teddy bear, pillow, and bed.
Spatial Story	The teddy bear is beside the pillow.
	The pillow is under the bed.
	The barrel, trash can, and basket.
Comparative Story	The barrel is smaller than the trash can.
	The trash can could fit inside of the basket.

Comparative Story Test Assertions

Information	Order	Validity	
Prem.	Orig.	Valid	The barrel is littler than the trash can.
		Invalid	The trash can could enclose the basket.
	Rev.	Valid	The basket could enclose the trash can.
		Invalid	The trash can is littler than the barrel.
Infer.	Orig.	Valid	The barrel could be inserted in the basket.
		Invalid	The barrel could enclose the basket.
	Rev.	Valid	The basket could enclose the barrel.
		Invalid	The basket could be inserted in the barrel.

Table 2: Mean Proportions of Valid (Upper Panel) and Invalid (Lower Panel) Test Assertions Correctly Verified as a Function of Instructions, Grade, Information, and Story

Valid

		Fifth Grade			Eleventh Grade		
Instructions	Information	Spatial	Comparative	Total	Spatial	Comparative	Total
Prompt	Premise	.62	.68*	.65*	.72*	.79*	.75*
	Inference	.58	.59	.58	.75*	.73*	.74*
Meaning	Premise	.71*	.81*	.76*	.81*	.75*	.78*
	Inference	.64	.52	.58	.73*	.79*	.76*
Verbatim	Premise	.70*	.75*	.72*	.75*	.62	.68*
	Inference	.35	.75*	.55	.79*	.79*	.79*

Valid

		Fifth Grade			Eleventh Grade		
Instructions	Information	Spatial	Comparative	Total	Spatial	Comparative	Total
Prompt	Premise	.73*	.77*	.75*	.83*	.77*	.80*
	Inference	.73*	.79*	.76*	.73*	.69*	.71*
Meaning	Premise	.68*	.66*	.67*	.87*	.81*	.84*
	Inference	.73*	.75*	.74*	.69*	.73*	.71*
Verbatim	Premise	.69*	.68*	.68*	.81*	.79*	.80*
	Inference	.60	.63*	.61*	.77*	.75*	.76*

* Mean proportion significantly above chance, or .50 (α = .05)

Table 3: Mean Proportions of Invalid Test Assertions Correctly Verified as a Function of Grade, Order, and Information

	Fifth Grade			Eleventh Grade		
Order	Premise	Inference	Total	Premise	Inference	Total
Original	.63	.67	.65	.83	.74	.78
Reversed	.78	.74	.76	.80	.70	.75

Table 4: Mean Proportion of Test Assertions Correctly Verified on Immediate Test (Upper Panel) and Delayed Test (Lower Panel) as a Function of Grade, Validity, Information, and Story

Comprehension

		Fifth Grade			Eleventh Grade		
Information	Validity	Spatial	Comparative	Total	Spatial	Comparative	Total
Premise	Valid	.91	.85	.88	.90	.96	.93
	Invalid	.79	.88	.83	.98	.92	.95
Inference	Valid	.69	.75	.72	.83	.90	.86
	Invalid	.90	.73	.81	.90	.83	.86

Memory

		Fifth Grade			Eleventh Grade		
Information	Validity	Spatial	Comparative	Total	Spatial	Comparative	Total
Premise	Valid	.68	.75	.71	.76	.72	.74
	Invalid	.70	.70	.70	.84	.79	.81
Inference	Valid	.53	.61	.57	.76	.77	.76
	Invalid	.69	.72	.70	.73	.71	.72

Table 5: Mean Proportions of Comparative Test Assertions Correctly Verified on Immediate Test (Upper Panel) and Delayed Test (Lower Panel) as a Function of Markedness, Grade, Validity, and Information

Comprehension

		Fifth Grade			Eleventh Grade		
Markedness	Validity	Premise	Inference	Total	Premise	Inference	Total
Unmarked	Valid	.85	.69	.77	1.00	.83	.91
	Invalid	.90	.69	.79	.96	.83	.89
Marked	Valid	.81	.69	.75	.92	.96	.94
	Invalid	.71	.77	.74	.87	.83	.85

Memory

		Fifth Grade			Eleventh Grade		
Markedness	Validity	Premise	Inference	Total	Premise	Inference	Total
Unmarked	Valid	.84	.75	.79	.83	.89	.86
	Invalid	.81	.85	.83	.92	.81	.86
Marked	Valid	.64	.50	.57	.60	.65	.62
	Invalid	.63	.63	.63	.68	.63	.65

Figure 1: Mean Proportions of Valid (Upper Panel) and Invalid (Lower Panel) Test Assertions Correctly Verified as a Function of Grade and Information.