

On the reliability and validity of children's metamemory

BETH E. KURTZ, MOLLY K. REID, and JOHN G. BORKOWSKI
University of Notre Dame, Notre Dame, Indiana 46556

and

JOHN C. CAVANAUGH
Bowling Green State University, Bowling Green, Ohio 43403

Test-retest reliability, internal consistency, and construct validity of a metamemory battery were assessed. Fifty-five second-graders were given an extensive metamemory battery on two occasions separated by a 6-week interval. Test-retest correlations ranged from .29 to .49 for individual subtests, with a correlation of .67 for the composite battery. The children were then given training on the use of an interrogative-elaboration strategy to aid paired associate learning. Following three training sessions, tests for strategy maintenance and generalization were administered. Correlations relating strategy use with metamemory were significant for the transfer tasks, even when general knowledge was partialled out. The concept of metamemory appears useful in explaining production deficiencies and insightful strategy invention in young children.

Metamemory refers to self-knowledge about memory processes (Brown, 1978; Flavell & Wellman, 1977). Recent studies have examined the construct validity of metamemory in terms of its relationship to a variety of strategic behaviors (Cavanaugh & Borkowski, 1980; Kendall, Borkowski, & Cavanaugh, 1980; Beuhring, Note 1). While significant correlations between memory knowledge and memory processes or recall accuracy have been found, the modest size of these relationships is not strongly supportive of extant metamemory theory. The low to moderate correlations may be due as much to weaknesses in the instruments measuring metamemory as to the theoretical invalidity of the concept (Cavanaugh & Borkowski, 1980).

The present studies attempted to measure the test-retest reliability, internal consistency, and construct validity of a new metamemory battery. In Experiment 1, 55 second-graders were given a seven-item metamemory battery on two occasions, with testing sessions separated by a 6-week interval. The metamemory test was composed of four items from the Kreutzer, Leonard, and Flavell (1975) scale, a memory monitoring task (Levin, Yussen, DeRose, & Pressley, 1977), and two new subtests assessing children's knowledge about input/output processing strategies. Experi-

ment 2 examined the role of metamemorial knowledge as a predictor of strategy transfer. That is, the second-graders, used originally to determine reliability, were trained on a four-step interrogative-elaboration strategy as an aid to paired associate learning. The elaboration strategy was not explicitly associated with any of the metamemory subtests. Kestner and Borkowski (1979) employed similar task and training procedures and found recall to be significantly related to the quantity and quality of elaborations used at transfer.

Borkowski (in press) has argued that metamemory/strategy-use relationships will be found only in limited contexts, such as on novel tasks or transfer tests. We hypothesized that metamemory, as indexed by our battery, would predict strategic behavior on paired associate tasks following the training of the elaboration strategy and would show stronger relationships than a measure of general intelligence. That is, young children given prolonged, thorough strategy training should be differentiated on transfer tests on the basis of prior sophistication in metamemorial knowledge.

EXPERIMENT 1

Method

Subjects. Fifty-five second-grade children (mean age = 7.7 years) from two parochial schools in South Bend, Indiana, served as subjects. The sample consisted of 26 males and 29 females; approximately 96% were Caucasian.

Materials. Four metamemory subtests were selected from the metamemory battery designed by Kreutzer et al. (1975): story list, preparation object, retrieval event, and rote paraphrase. These subtests have been shown to be the most valid and sensitive indicators of metamemory within the Kreutzer inter-

This research was supported by NIE Grant G-81-0134. The first and second authors were supported by NIH Training Grant HD-07184. Reprint requests, as well as information on test construction and scoring procedures, can be obtained from John G. Borkowski, Department of Psychology, University of Notre Dame, Notre Dame, Indiana 46556.

view questionnaire for children between 6 and 9 years of age (Cavanaugh & Borkowski, 1980). Three other subtests, memory monitoring, interest categories, and serial recall, differed from the first four by being active measures of metamemory. That is, the first four subtests consisted of a verbal exchange regarding the actions of a second, hypothetical child, whereas the final three encouraged the child to actively engage in tasks requiring metamemorial knowledge.

"Story list" assessed knowledge of the efficacy of elaboration. Children were shown eight pictures (man, bed, tie, shoes, table, dog, hat, car) and asked if the items would be easier to recall after hearing them named or after hearing a story that included the eight. "Preparation object" measured children's planful behavior in preparing for future retrieval. The situation involved a child who wanted to skate after school on the next day. Children were asked what they might do tonight, if they were that child, in order to remember to take their skates to school the next day. "Retrieval event" measured systematic memory search for a past event. Children were again placed in a hypothetical situation, this time attempting to determine the age of a dog that had been given to them as a Christmas present, and were asked how they might go about remembering which Christmas they had received the dog. "Rote paraphrase" determined if children acknowledged the relative ease of gist recall over rote recall. "Memory monitoring," adapted from Levin et al. (1977), assessed knowledge of short-term memory capacity, especially the ability to adjust recall estimates based on a prior recall experience. A stack of 15 cards was used to elicit the first estimate of the number that might be remembered following a single study trial; then a second set of 15 pictures was displayed at 3-sec intervals, and recall was measured. A third set of cards, with a single new picture on top, was used to elicit a second estimate of recall; no study or recall trials were given. The scoring measure used was $|P - A|/A$, where P = number predicted in the second estimate and A = actual number recalled. "Interest categories" concerned awareness of the role of familiarity of subject matter in the allotment of study time in a memory task, and "serial recall" measured the ability to match output demands with input strategies.

Procedure. Children were tested individually in an area near their classroom. The seven subtests were given in a single 30-min session, without time restraints. Following a 6-week interval, the battery was readministered. Reliability of independent scorers was found to be .96 for the battery when averaged over both sessions.

Results

Scores obtained on each of the seven metamemory subtests were converted to z scores in order to give equal weighting to the subtests. Pearson product-moment correlations of the resulting z scores from the two testing sessions are shown in Table 1. With the exception of

Table 2
Intersubtest Correlations and Correlations of Each Subtest With the Sum of All Others

Sub-test	Subtest					
	1	2	3	4	5	6
2	.13					
3	-.03	.51†				
4	.25	.35*	.43*			
5	.04	-.02	.19	.20		
6	.35*	.10	.21	.37*	.22	
7	.23	.34*	.45†	.55†	.19	.40*

Note—1 = story list; 2 = preparation object; 3 = retrieval event; 4 = rote paraphrase; 5 = memory monitoring; 6 = interest categories; 7 = $\sum_{i=1}^6$. * $p < .01$. † $p < .001$.

serial recall, all subtests had significant test-retest correlations, ranging from $r = .29$ ($p < .01$) for preparation object to $r = .49$ ($p < .001$) for retrieval event. Serial recall was excluded from the battery at this juncture because of its poor reliability. Composite metamemory scores for each subject were obtained by summing z scores from the remaining six subtests. The test-retest correlation of the composite battery was $r = .67$ ($p < .001$).

In order to assess the internal consistency of the battery, a correlation matrix was produced on data from the first session. This matrix, shown in Table 2, reflects the relationships between pairs of subtests, as well as the correlation of each subtest with the sum of the other five subtests. Five of the 15 correlations between pairs of subtests were significant, and the arithmetic mean of the 15 correlations was .22. Of the six subtests, only story list and memory monitoring failed to correlate significantly with the sum of the other five subtests. The other four correlations were moderate, ranging from $r = .34$ ($p < .01$) for preparation object to $r = .55$ ($p < .001$) for rote paraphrase. A split-half reliability measure of data from the first testing session (i.e., the first three subtests correlated with the last three) yielded $r = .42$ ($p < .01$). The correlation between summed scores from the first, third, and fifth subtests and summed scores from the second, fourth, and sixth subtests was $r = .57$ ($p < .001$).

EXPERIMENT 2

Method

Subjects, Design, and Materials. The 55 second-graders from Experiment 1 served as subjects for Experiment 2. Following Experiment 1, children were assigned to experimental ($N = 43$) and control ($N = 12$) groups, with the stipulation that metamemory scores of the control group be representative of scores in the entire sample. The experimental group received training on an interrogative-elaboration strategy during three training sessions (cf. Kestner & Borkowski, 1979). Maintenance and generalization of the strategy were tested during a final testing session 1 week after the end of training. All children received the vocabulary subtest of the WISC.

Training and maintenance tasks were conducted with pictures mounted on 7.5 x 12.5 cm index cards. Two pictures per

Table 1
Test-Retest Correlations for Individual Subtests and the Composite Battery

Subtest	r	p
Story List	.37	.002
Preparation Object	.29	.01
Retrieval Event	.49	.001
Rote Paraphrase	.40	.001
Memory Monitoring	.30	.014
Interest Categories	.33	.01
Serial Recall	.13	.16
$\sum_{i=1}^7$.67	.001

card were used for the study trials. One picture of each pair served as a cue for recall on test trials. For the generalization task, three pictures were mounted on 13 x 18 cm index cards for study trials. One picture again functioned as a cue for recall. Each training session contained 8 pairs, the maintenance task contained 16 pairs, and the generalization task contained eight picture triads. Identical materials were used for the experimental and the control groups.

Procedure. The experimental group was taught a four-step elaboration strategy modeled after a procedure developed by Kestner and Borkowski (1979) for use with paired associates: (1) associating the to-be-remembered pair together, (2) asking "why" the two unrelated items are together, (3) elaborating on characteristics of the items, and (4) answering the "why" question using elaborations generated in Step 3. Three training sessions were conducted at 1-week intervals, using a familiar experimenter for the first two sessions and a new experimenter for the third session. Initially, the four-step process was demonstrated by the experimenter, and then it was repeated by each child with steadily diminishing experimenter prompts. Children assigned to the control group spent an equivalent amount of time with the experimenter and the materials, but they were given no strategy instructions. After viewing the pictures at 15-sec intervals, the children were tested for recall.

Maintenance and generalization were assessed by a new experimenter 1 week after the final training session. Children were not reminded of the strategy. In the generalization task, triads of pictures were shown at 20-sec intervals. After eight triads had been viewed, recall of the last two pictures in each set was requested, using as a picture prompt the first member of each triad. The order of presentation on the test trial was random. Following recall, children were tested with a probe task in order to detect the quality of elaborations used in studying the original picture pairs. Scoring was similar to the procedure used by Kestner and Borkowski (1979). The maintenance task was administered in approximately the same fashion as the generalization test. Following a study-test trial, probe questions were used to measure strategy use for each item; responses to the probes were again scored in terms of the amount of elaboration and association (cf. Kestner & Borkowski, 1979).

Results

Table 3 shows the data for both recall and strategy use on the maintenance and generalization tasks for trained and untrained conditions. All experimental-control differences were significant, except for strategy use on the word-triad task. A visual analysis of the transfer data revealed striking variability in strategy use within the experimental group. Two distinct subgroups emerged following training: strategy users and strategy nonusers. Using 12 points (from 64 possible in the maintenance task) as a cutoff point, 19 of the children

were classified as nonusers, with a mean strategy score of .32 and an SD of 1.00. Twenty-four children used the strategy, with a range of 15 to 64 points (mean = 50.71, SD = 13.42). Using 12 of 56 possible points as a cutoff point for the generalization task, 24 children qualified as nonusers, with a mean strategy score of zero, whereas 19 children were classified as strategy users (mean = 47.26, SD = 10.21).

Point biserial correlations of strategy use and recall were significant for both the maintenance ($r_{pb} = .65$, $p < .001$) and generalization tasks ($r_{pb} = .58$, $p < .001$). The mean number recalled by strategy users and nonusers were compared and found to be significantly different in favor of the user group for maintenance [$t(41) = 6.00$, $p < .001$] and generalization [$t(41) = 4.17$, $p < .001$].

Point biserial correlations of strategy use and metamemory were significant both in the maintenance task ($r_{pb} = .39$, $p < .01$) and in the generalization task ($r_{pb} = .26$, $p < .05$). Performance on the WISC subtest and metamemory battery correlated significantly ($r = .46$, $p < .001$). However, when partial correlations were used to remove the effect of WISC scores from the metamemory/strategy-use correlation, the maintenance task correlation remained significant ($r = .34$, $p < .05$); for the generalization task, the correlation between metamemory and strategy use became nonsignificant ($r = .21$). Interestingly, general knowledge, as measured on the WISC subtest, failed to predict strategy use in either transfer setting.

DISCUSSION

Previous studies examining metamemory in children have generally not been concerned with the issue of reliability. Data from Experiment 1 indicate that many individual subtests that have previously been used to assess metamemory have a significant but low level of reliability. This should cause hesitation before researchers adopt a single, brief assessment procedure as the operational definition of metamemory. However, when individual subtest scores were pooled to form a composite battery score, a more acceptable level of reliability was obtained. Increasing the number of items in each subtest, clarifications in instructions and scoring, and revisions in the battery to obtain a broader representation of memory phenomena will likely lead to even higher reliability and, perhaps, greater internal consistency.

Our results corroborated those of Kendall et al. (1980), who found that metamemory predicted the use of an elaboration strategy on transfer tests by retarded children. The present results strengthen the construct validity of metamemory by demonstrating that a group of normal children given extensive strategy training profit differentially depending on depth and extent of prior knowledge about memory processes. It should be emphasized that metamemory predicted strategy use at maintenance even when an estimate of intelligence was partialled out. In all transfer contexts, metamemory was a better predictor of memory processing than general intelligence was.

Why is the concept of metamemory important to theory construction in cognitive development and intelligence? One possibility is that metamemory serves as a causal concept in explaining the production deficiencies of young children. That is, production deficiencies may reflect an inadequate understand-

Table 3
Experimental-Control Group Differences
on Transfer Tasks

Group	Maintenance		Generalization	
	Recall	Strategy Use	Recall	Strategy Use
Experimental	6.88	29.12	5.19	22.33
Control	3.08	11.83	2.58	12.83
p <	.01	.025	.01	.16

Note— p = two-tailed probability.

ing of the function and importance of trained strategies. This interpretation squares well with studies on the role of feedback following strategy training that generally show dramatic improvement in strategy durability on transfer tests (cf. Borkowski, in press). In a similar vein, our results suggest that children who have high levels of metamemorial knowledge engage more readily in strategic behavior than do their low-metamemory peers who have been given equal training on an identical strategy. However, before pursuing the issue of construct validity in greater earnest, more refined, wide-ranging measurement operations seem necessary in order to produce indexes of metamemory with high reliability.

REFERENCE NOTE

1. Beuhring, T. *Developmental differences in elaboration: A metamemory explanation*. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1979.

REFERENCES

- BORKOWSKI, J. G. Signs of intelligence: Strategy generalization and metacognition. In S. Yussen (Ed.), *The development of reflection in children*. New York: Academic Press, in press.
- BROWN, A. L. Knowing when, where, and how to remember: A problem of metacognition. In R. Glasser (Ed.), *Advances in instructional psychology* (Vol. 1). Hillsdale, N.J.: Erlbaum, 1978.
- FLAVELL, J. H., & WELLMAN, H. M. Metamemory. In R. V. Kail & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition*. Hillsdale, N.J.: Erlbaum, 1977.
- CAVANAUGH, J. C., & BORKOWSKI, J. G. Searching for metamemory-memory connection: A developmental study. *Developmental Psychology*, 1980, 16, 441-453.
- KENDALL, C. R., BORKOWSKI, J. G., & CAVANAUGH, J. C. Maintenance and generalization of an interrogative strategy by EMR children. *Intelligence*, 1980, 4, 255-270.
- KESTNER, J., & BORKOWSKI, J. G. Children's maintenance and generalization of an interrogative learning strategy. *Child Development*, 1979, 50, 485-494.
- KREUTZER, M. A., LEONARD, C., & FLAVELL, J. H. An interview study of children's knowledge about memory. *Monographs of the Society for Research in Child Development*, 1975, 40(Serial No. 159).
- LEVIN, J. R., YUSSEN, S. R., DEROSE, T. M., & PRESSLEY, M. Developmental changes in assessing recall and recognition memory capacity. *Developmental Psychology*, 1977, 13, 608-615.

(Received for publication December 17, 1981.)