Bulletin of the Psychonomic Society
1984, 22 (6), 493-496

Evidence for long-term planning in children's hypothesis testing

JOAN H. CANTOR and CHARLES C. SPIKER
University of Iowa, Iowa City, Iowa

Children in kindergarten through third grade were given a series of specially designed discrimination tasks in which the stimuli were toy animals rather than conventional multidimensional (e.g., color-form-size) compounds. The purpose was to assess the hypothesis-testing skills of young children in a situation in which they would not tend to perseverate on irrelevant features of the stimuli. Their ability to make strategic decisions was studied across tasks, in terms of measures of their effective use of immediate feedback information, as well as in terms of measures of their long-term planning across trials. Although age differences were found, children at all age levels learned the tasks very rapidly and showed a high degree of sophistication in hypothesis testing. Improvement was shown across training tasks, particularly in long-term planning. The results demonstrate for the first time clear evidence of such long-term planning among kindergarten children.

During the past decade, considerable research effort has been directed toward determining the extent to which children in the early elementary years systematically test hypotheses as a means of solving discrimination problems (see Spiker & Cantor, 1983, and Tumblin & Gholson, 1981, for reviews). Beginning at the second-grade level, children show evidence of testing hypotheses under a wide variety of conditions (e.g., Kemler, 1978; Phillips & Levine, 1975). Children at the kindergarten level, however, appear to test hypotheses under more restricted conditions, for example, in highly simplified tasks (e.g., Cantor & Spiker, 1978), following various types of preliminary training (e.g., Spiker & Cantor, 1979), or in the context of story-and-game problems (Kemler, 1978). Furthermore, results to date have failed to provide clear evidence that children under 7 years of age show long-term planning across trials in using hypothesis-testing techniques to move toward the solution of the problem. Kemler (1978) made comparisons across several age levels and concluded that kindergarten children, although effectively using current feedback information to select new hypotheses, did not remember which hypotheses they had disconfirmed on previous trials. Second-graders, on the other hand, did show evidence of memory for previously disconfirmed hypotheses.

The purpose of the present study was to investigate hypothesis testing in children in kindergarten through third grade under conditions that might provide evidence of such long-term planning even among the youngest children. For this purpose, an attempt was made to reduce the difficulties that young children frequently have in solving conventional simultaneous discrimination problems. In these tasks, the subject must learn to select on the basis of a single dimensional value (e.g., red) when pairs of multidimensional compounds (e.g., large red square and small blue circle) are presented. Investigators have repeatedly reported that many children, particularly in kindergarten and first grade, are hampered in such tasks by a tendency to limit their hypotheses to the values of a single irrelevant stimulus dimension (e.g., Kemler, 1978; Spiker & Cantor, 1983).

In an attempt to reduce such perseverative responses and thereby expose more sophisticated hypothesis testing, we developed a new task in which the stimuli are familiar objects rather than the usual dimensional compounds. Eight toy animals serve as potential solutions in a two-choice task in which the child attempts to discover which one of the eight animals is "special." On each trial, the eight animals are presented in two sets of four, and the child chooses the set he or she thinks contains the special animal. Using feedback information that is provided across trials, the child can determine which animal (e.g., lion) always appears in the correct set, and is therefore the solution. The animals are not paired to form relevant and irrelevant dimensions comparable to those in the multidimensional tasks, but are regrouped from trial to trial so that all except for the "special" animal appear in both correct and incorrect sets. In other respects, the task is analagous to a four-dimensional simultaneous discrimination problem with one relevant and three irrelevant dimensions.

The child's hypotheses were determined through the use of posttrial introtact probes, that is, requests for the child's current best guess at the solution, given at the end of each

The authors are indebted to Philip Hershner, Gayle Klouda, and Kay Sabin for collecting and tabulating the data. We are also grateful to David L. Cronin, Richard E. Taylor, and Donald J. Tvedte in the Iowa City School District for their excellent cooperation. Funds for the purchase of computer time were made available by the Graduate College of the University of Iowa. Requests for reprints should be addressed to Joan H. Cantor, Department of Psychology, University of Iowa, Iowa City, IA 52242.

trial with feedback information in full view. Hypothesis testing in kindergarten children has been shown to be better with posttrial probes than with the more conventional pretrial probes given prior to the actual choice response (Cantor & Spiker, 1982).

METHOD

Subjects

The subjects were 118 children, 28 in kindergarten and 30 each in first, second, and third grades in a predominantly upper middle-class school system, who participated on a voluntary basis after their parents had returned consent slips. The mean and range of chronological age for the children in kindergarten through third grade were 70.68 (64-81) months, 83.90 (76-100) months, 96.23 (89-107) months, and 108.50 (98-120) months, respectively. The data for an additional 12 children were eliminated, 4 due to experimenter errors (1 at each grade level), 4 due to lack of cooperation (all at kindergarten level), and 4 for failure to follow instructions (2 at kindergarten and 2 at second-grade levels).

Stimulus Materials, Apparatus, and Tasks

Stimuli. The stimuli were eight small plastic animals—bear, cow, elephant, giraffe, horse, lion, pig, and turtle. The animals ranged in height from about 3 cm for the pig to approximately 7 cm for the giraffe. Each was of such a size that it could be mounted on top of a 2.5 × 5 mm clear Plexiglass base.

Apparatus. The apparatus, painted black, consisted of a horizontal platform (56 cm wide × 44 cm deep) mounted on a turntable bisected lengthwise by a vertical partition (32 cm high). The stimuli were presented on the platform in front of the partition. A red card (26.0 × 16.0 cm) was attached to the right side of the platform, and a blue card of the same size was attached to the left side. On each trial, half of the animals were placed on the red card, and the other half on the blue card. Between trials, the platform was rotated, removing the animals from the subject's view so that they could be regrouped for the next trial.

Tasks. Each child received three preliminary tasks followed by one criterion task. Four of the eight animals were used in the first pretraining task (Task P1), the remaining four animals were used in the second pretraining task (Task P2), six animals were used in the third pretraining task (Task P3), and all eight animals were used in the criterion task (Task C).

For purposes of counterbalancing, the animals were arbitrarily divided into two sets, Set A, consisting of the lion, cow, giraffe, and pig, and Set B, consisting of the turtle, bear, elephant, and horse. For successive tasks, the animal that constituted the solution was drawn alternately from Sets A and B, starting with Set A in one counterbalancing condition and Set B in the other counterbalancing condition.

On each learning trial, half of the animals were placed in a horizontal array on the red card, and the remaining half were placed on the blue card. The child attempted to choose the card (red or blue) that contained the "special animal," and was given feedback after each choice. The animals were regrouped following each trial in such a way that, beginning with any trial, the use of a maximally efficient focusing strategy (complete memory for all potential solutions that have been eliminated by prior feedback) would lead to solution by the end of two trials in the four-animal problems and by the end of three trials in the six- and eight-animal problems. The correct animal appeared equally often on the blue (left) and the red (right) cards within blocks of four trials in Tasks P1 and P2, within blocks of six trials in Task P3, and within blocks of eight trials in Task C. In addition, each animal appeared equally often in the various positions within the array on the red and blue cards.

Procedure

Pretraining. The experimenter presented the four animals to be used in Task P1 and named each in turn. The experimenter explained that he or she was thinking about one special animal, that any one of the four animals might be the special one, and that the child's task was to figure out which animal was special. The child was then asked, "Which one do you think might be the special animal?" This initial introtact probe was given in each task, and the child's response was used to determine the assignment of the positive cue. In all pretraining and criterion tasks, one of the animals not named as the child's first hypothesis was assigned as the positive cue. In all tasks except for Task P2, negative feedback was given following the child's first-choice response. In Task P2, positive feedback was given on the first trial in order to minimize the likelihood that the child would assume that his or her first choice would always be incorrect. Since young children frequently conclude that the experimenter changes the solution during the course of a learning task, a procedure was routinely adopted of hiding a duplicate of the positive cue in a small metal file box. The child was told that "an animal exactly like the special one is in this box." The box was left in the child's view, and at the end of the task, the child was allowed to look inside to verify the solution.

The child was told that two animals would appear on each card, and that the special animal would be on the red card sometimes and on the blue card sometimes. The experimenter then said, "You tell me each time whether you think the blue card or the red card has the special animal on it." After the child had made a choice, the experimenter placed a white cardboard strip along the full length of the correct card and said, "The (red) (blue) card has the special animal on it this time.... After I put the white strip down each time, you think about it and tell me which animal you think might be the special one. Then I'll give you your next turn." Thus, the child's hypotheses were obtained using posttrial introtact probes, that is, probes at the end of each trial with full feedback information in view. Because of the relatively long time interval between the statement of hypothesis and the next choice of the red or blue card, many children tended to make inconsistent responses. In an attempt to teach the child to be response consistent, the child was required, in the preliminary tasks, to choose in accord with the previously stated hypothesis. During the criterion task, the child was free to make inconsistent responses.

In Task P1, the child was brought to a criterion of 4/4 correct choices with concomitant correct hypotheses. If the criterion had not been met within six trials, the experimenter gave the correct solution to the child, and additional trials were given until the criterion had been met.

In Task P2, the new animals were introduced to the child, and, with the exception noted above, the same procedures were followed as for Task P1. In Task P3, six animals were used; the child was told that there would be three animals on each card in this game, and that it might take a little longer to figure out which animal was special. In other respects, the procedures were the same as in the first two tasks except that eight trials were given before the experimenter provided the child with the solution.

Criterion task. Before beginning the criterion task, the children were told that all eight animals would be used, and that four animals would appear on each card. The experimenter then said, "I'll give you lots of turns in this game, and I won't help you this time. Let's see if you can figure out which animal is special all by yourself." The learning criterion was the same as that used during pretraining, and the maximum number of trials was 24.

Response Measures

Discrimination-learning performance on the preliminary and criterion tasks is measured by determining the trial of last error for each child. Six additional measures that are derived from the child's responses to the introtact probes are used to assess the extent of his or her strategic skills. *Response consistency* is the proportion of trials in which the child actually selects the stimulus compound containing the value named in the current hypothesis. Adherence to the *win-stay* rule is measured by the proportion of trials following positive feedback in which the child maintains his previous hypothesis. Adherence to the *lose-shift* rule is measured in terms of the proportion of trials following negative feedback in which the child selects a new hypothesis. Adherence to the *local-consistency* rule is measured by the proportion of trials following negative feedback in which the child selects as a new hypothesis a value from the currently displayed positive compound. Adherence to the *valid-hypothesis* rule is based on only those trials in which the child changes hypotheses. The probability is the proportion of these trials for which the "new" hypothesis is one that has not been previously tested and disconfirmed. *Cumulative consistency* is a measure of the child's memory for locally consistent hypotheses and is closely related to a measure labeled "stimulus memory" by Kemler (1978). It is the proportion of trials following negative feedback in which the child selects as a new hypothesis a value that is cumulatively locally consistent with feedback information on the two most recent trials (i.e, the current and immediately preceding trials).

RESULTS AND DISCUSSION

Discrimination Learning

The means of the trial of last error (TLE) in the criterion task for the kindergarten, first-grade, second-grade, and third-grade children were 8.64, 7.03, 3.80, and 3.50, respectively. These grade differences were significant [$F(3,114) = 4.32$, $p = .006$]. The proportions of learners in the four grades were .79, .87, .97, and 1.00, respectively. Again the grade differences were significant [$\chi^2(3) = 9.78$, $p < .025$]. Follow-up analyses of both response measures indicated that there was no difference between kindergarten and first grade or between second and third grade, but that the combined second and third grades were significantly superior to the combined kindergarten and first grade ($p < .005$ in both cases).

In spite of the reliability of these age effects, the differences were relatively small. On the average, the kindergarteners required only about five more trials than did the third-graders. Moreover, approximately 80% of the kindergarten children met criterion in the 24 trials allotted. These results compare quite favorably with those reported by Kemler (1978) for her kindergarten children. In her task, the children had to solve for the secret article of clothing that identified "Amy," one of a pair of pictures of twin girls. The kindergarten children received a task with four dimensions (e.g., belts, ribbons, eyeglasses, necklaces), and therefore eight possible solutions, so in that respect her task was quite comparable to ours. Kemler reported a mean of 15.1 for the TLE, whereas the mean TLE for our kindergarten children was only 8.64. Although the present criterion task had eight potential solutions, it was only slightly more difficult for kindergarteners than the conventional task in an earlier experiment (Cantor & Spiker, 1982) that had only four potential solutions (mean TLE = 5.78). It seems quite likely that the high level of discrimination performance is a direct result of the elimination of the dimensional perseveration usually found in more conventional tasks.

Indices of Strategy

The effects of practice. Although task complexity (i.e., number of potential solutions) is confounded with the number of problems previously solved, it may be worthwhile to examine the changes in the various components of strategy that occur with practice. Since the indices of strategy are all proportions, they may be compared across the preliminary and criterion tasks.

The local-consistency and lose-shift indices were virtually perfect for all age levels in all of the tasks and were not further analyzed. The high levels of these measures replicate earlier results (Cantor & Spiker, 1982; Kemler, 1978) and are consistent with our contention that posttrial probes constrain the child to be locally consistent and thereby to adhere to the lose-shift rule.

The means for the remaining indices of strategy are presented in Table 1 for preliminary Tasks P1 and P3, as well as for the criterion task. The response-consistency measure is not included for the preliminary tasks, since the children were coached to be response consistent during these tasks. Task P2 is not included because the procedures were somewhat different for this task.

The measures shown in Table 1 were subjected to a multivariate analysis of variance (MANOVA), with grade as a between-subjects factor, task as a within-subject factor, and the three indices as dependent measures. Only the main effects of grade and task were significant, [$F(9,273) = 2.82$, $P < .01$, and $F(6,109) = 5.90$, $p < .001$, respectively]. Follow-up tests revealed that the combination of Grades 2 and 3 was significantly superior to the combination of Grades K (kindergarten) and 1 for all three measures (all $ps < .001$). Follow-up tests for the task effect showed significant improvement across tasks for all three dependent measures (all $ps < .001$), despite the increase in task complexity.

An additional MANOVA was conducted for the criterion task alone, with grade as a between-subjects factor and including response consistency as a fourth dependent measure. The main effect of grade was significant [$F(12,294) = 1.87$, $p < .05$]. Follow-up tests again showed superiority of the combined Grades 2 and 3 over the combined Grades K and 1 for all of the response measures (all $ps < .05$).

Short-term vs. long-term efficiency. The local-consistency, lose-shift, and win-stay measures provide evidence of the child's ability to make short-term use of current feedback information in testing hypotheses. Even the youngest children were perfect on the first two of these measures, and the win-stay probabilities were also quite high in the criterion task. Thus, the results show that short-term efficiency was very good for all grades, with only the win-stay measure needing any improvement with age. These high probabilities for short-term efficiency are entirely comparable to those reported by Kemler (1978) for her kindergarten children. As noted earlier, the excellent performance of the younger children on these three indices was expected because of the use of posttrial probes.

Table 1
Means for the Indices of Strategy Across Preliminary and Criterion Tasks for the Four Grade Groups

	Task P1			Task P3			Criterion Task			
Grade	W-S	V-H	C-C	W-S	V-H	C-C	W-S	V-H	C-C	R-C
Kindergarten	.70	.84	.64	.82	.86	.65	.83	.80	.72	.87
First Grade	.75	.76	.55	.89	.90	.64	.87	.88	.77	.89
Second Grade	.85	.87	.75	.93	.94	.80	.96	.96	.85	.93
Third Grade	.87	.88	.75	.89	.94	.78	.98	.97	.87	.96

Note—W-S = win-stay; V-H = valid hypothesis; C-C = cumulative consistency; R-C = response consistency.

Table 2
Proportions of Children with Errorless Performance on the Criterion Task in the Four Age Groups

	Grade			
	Kindergarten	First	Second	Third
Short-Term Efficiency				
Local Consistency	1.00	1.00	1.00	1.00
Lose-Shift	1.00	1.00	1.00	1.00
Win-Stay	.68	.63	.93	.93
Long-Term Efficiency				
Valid Hypothesis	.64	.67	.87	.93
Cumulative Consistency	.50	.50	.60	.93
Response Consistency	.50	.60	.67	.83
All Indices	.39	.47	.53	.63

One of the major questions in the present study is whether younger children also demonstrate long-term efficiency by using information accumulated across trials to select new hypotheses. Two measures of long-term planning, the valid-hypothesis probability and cumulative consistency, were both quite high even for the younger children. Nevertheless, a strong case for long-term efficiency requires a demonstration that performance is better than would be expected on the basis of short-term efficiency alone. A test suggested by Kemler (1978) was used for this purpose.

This test is based on an analysis of performance on cumulative consistency, which it will be recalled is the proportion of trials following negative feedback in which the child's new hypothesis is cumulatively consistent with feedback information on the most recent pair of trials. Given the sequences of stimuli used, the probability of being cumulatively consistent on the basis of short-term strategy alone is .50, since the positive compounds on successive trials in the criterion task always have two of four animals in common. It can be seen in Table 1 that all age groups had values for cumulative consistency considerably higher than .50, and that even the kindergarten children had a value of .72. All age groups were significantly higher than .5 in the criterion task, with z values of 4.57, 5.84, 7.78, and 8.02 for kindergarten through third grade, respectively (all ps $<$ .0001). Comparable tests made for cumulative consistency in Tasks P1 and P3 indicate that all were significantly higher than .5 (all ps $<$ .01), except for the first-graders in Task P1. Thus, these tests provide clear evidence of the presence of a long-term strategy even in the youngest age group in this study. These results confirm Kemler's (1978) findings of evidence for long-term efficiency in second-graders, but differ from hers in that they provide such evidence even in kindergarten children. It seems likely that the superiority in long-term planning demonstrated here, relative to that reported by Kemler, is attributable to the reduction in perseveration in our dimensionally unstructured tasks.

Errorless strategies. Another way to evaluate the hypothesis-testing performances of the children is to examine the proportion of children in each age group who made no errors on each of the components of strategy. Table 2 gives the proportion of children in each grade who performed without error with respect to each of the six indices of strategy and the proportion in each grade who were perfect on all six indices. As can be seen, the majority of children were perfect on most of the measures. The children seemed to have the most difficulty in being response consistent and in maintaining cumulative consistency.

GENERAL CONCLUSIONS

One of the most important results in the present study was the demonstration of long-term planning in kindergarten and first-grade children. In fact, these children showed a high level of performance in nearly all components of strategy. These findings strongly support our original hypothesis that young children would reveal far more sophisticated problem-solving strategies in tasks in which the children do not perseverate in testing hypotheses about irrelevant features of the stimuli. In conventional multidimensional tasks, children of these ages have previously shown comparably high levels of performance only in highly simplified tasks (Spiker & Cantor, 1983). Older children still perform better than younger children, but their advantage has been considerably reduced in the tasks used here.

The data in Table 2 are especially descriptive of the problem-solving capabilities of children of these ages. The proportions of children who performed perfectly on all the measures of strategy, both short-term and long-term, are genuinely impressive. Furthermore, the proportions of children who obtained perfect cumulative consistency and valid-hypothesis probabilities indicate that sizable numbers of children at all grade levels showed truly outstanding performance.

REFERENCES

Cantor, J. H., & Spiker, C. C. (1978). The problem-solving strategies of kindergarten and first-grade children during discrimination learning. *Journal of Experimental Child Psychology*, **26**, 341-358.

Cantor, J. H., & Spiker, C. C. (1982). The effect of the temporal locus of the introtact probe on the hypothesis-testing strategies of kindergarten children. *Journal of Experimental Child Psychology*, **34**, 510-525.

Kemler, D. G. (1978). Patterns of hypothesis testing in children's discriminative learning: A study of the development of problem-solving strategies. *Developmental Psychology*, **14**, 653-673.

Phillips, S., & Levine, M. (1975). Probing for hypotheses with adults and children: Blank trials and introtacts. *Journal of Experimental Psychology: General*, **104**, 327-354.

Spiker, C. C., & Cantor, J. H. (1979). Factors affecting hypothesis testing in kindergarten children. *Journal of Experimental Child Psychology*, **28**, 230-248.

Spiker, C. C., & Cantor, J. H. (1983). Components in the hypothesis-testing strategies of young children. In T. Tighe & B. E. Shepp (Eds.), *Perception, cognition, and development: Interactional analyses*. Hillsdale, NJ: Erlbaum.

Tumblin, A., & Gholson, B. (1981). Hypothesis theory and the development of conceptual learning. *Psychological Bulletin*, **90**, 102-124.

(Manuscript received for publication June 18, 1984.)