ELSEVIER

Cognition 53 (1994) 129–153

COGNITION

The relationship between cognition and action: performance of children $3\frac{1}{2}$–7 years old on a Stroop-like day–night test

Cherie L. Gerstadt, Yoon Joo Hong, Adele Diamond*

Department of Psychology, University of Pennsylvania, 3815 Walnut Street, Philadelphia, PA 19104-6196, USA

(Received September 20, 1993, final version accepted April 26, 1994)

Abstract

One hundred and sixty children $3\frac{1}{2}$–7 years of age (10 M, 10 F at each 6-month interval) were tested on a task that requires inhibitory control of action plus learning and remembering two rules. They were asked to say "day" whenever a black card with the moon and stars appeared and to say "night" when shown a white card with a bright sun. Children <5 years had great difficulty. They started out performing well, but could not sustain this over the course of the 16-trial session. Response latency decreased from $3\frac{1}{2}$ to $4\frac{1}{2}$ years. Children <$4\frac{1}{2}$ years performed well when they took very long to respond. To test whether the requirement to learn and remember two rules alone was sufficient to cause children difficulty, 80 children $3\frac{1}{2}$–5 years old were tested on a control version of the task ("say 'day' to one abstract design and 'night' to another"). Even the youngest children performed at a high level. We conclude that the requirement to learn and remember two rules is not in itself sufficient to account for the poor performance of the younger children in the experimental condition.

1. Introduction

For over fifty years, the Stroop color-word task, created by J. Ridley Stroop (1935), has been studied with adults. In this task, color words (e.g., the words "blue" or "red") are printed in the ink of another color. Subjects are instructed to report the color of the ink rather than the word. This

* Corresponding author. E-mail diamond@cattell.psych.upenn.edu; 215 573-3892

SSDI 0010-0277(94)00632-U

requires that subjects inhibit their normal tendency when reading which is to attend to the words, ignoring the color of the ink. Some memory is required since subjects must remember that the task is to name the color of the ink, but the principal ability required is inhibition. Perret (1974) found that patients with damage to frontal cortex in the left hemisphere performed significantly worse on this task than patients with damage to other areas of the brain. This suggests that proper functioning of frontal cortex may be important for successful performance of the Stroop task. Subsequent studies, however, have not always replicated this finding (e.g., Stuss, Benson, Kaplan, Weir, & Della Malva, 1981).

Diamond (1988, 1990, 1991a) has hypothesized that frontal cortex is required whenever both memory (or sustained attention) and inhibition are needed. For example, consider Piaget's $A\bar{B}$ task (Piaget, 1954), which requires that subjects *keep in mind* where a toy was hidden on the present trial and *inhibit* their tendency to reach back to where they found the toy earlier (and, hence, where they were rewarded). Diamond and Goldman-Rakic have demonstrated that lesions of dorsolateral prefrontal cortex severely disrupt $A\bar{B}$ performance in adult monkeys (Diamond & Goldman-Rakic, 1989) and in infant monkeys (Diamond & Goldman-Rakic, 1986). Lesions to parietal cortex (Diamond & Goldman-Rakic, 1989) or to the hippocampal formation (Diamond, Zola-Morgan, & Squire, 1989) leave $A\bar{B}$ performance at comparable delays intact.

Several tasks on which Luria (1973) has found patients with extensive damage to frontal cortex to be impaired, such as the tapping test, require both memory and inhibition. On the tapping test, when the experimenter taps once the subject must tap twice, and when the experimenter taps twice the subject is to tap once. Here, the subject must remember both rules and inhibit the tendency to mirror the experimenter's actions. Luria found that adults with frontal cortex damage fail this test because they revert to doing what the experimenter does.

The present study examined the performance of children $3\frac{1}{2}$–7 years of age on a simplified version of the Stroop test. Our task contains a deck with two kinds of cards: the face of half the cards is white with a brightly colored sun, to which the subject is instructed to say "night". The face of the other cards is black with a moon and stars, to which the subject is instructed to say "day". Like the adult Stroop test, this Stroop-like day-night task requires subjects to inhibit a natural tendency to give a different verbal response. Unlike the adult Stroop task, however, which requires inhibition but little memory, our task taxes both memory and inhibition.

Since we hypothesized that younger children have difficulty remembering two rules while at the same time inhibiting a natural tendency, we predicted they would have difficulty with our day–night task. That is, we predicted that younger children would have a lower percentage of correct responses, and would need more time to formulate each response than older children.

One possibility is that younger children might perform poorly because they did not understand the instructions. We tried to minimize this possibility by giving each child preliminary training. We predicted that after children demonstrated that they understood what we were asking them to do by passing practice trials, even the youngest children would perform well on the initial trials. That is, we predicted that the younger children's problems would become more evident as a session progressed. Specifically, we predicted that the difference between performance early in a session versus later would be greater for younger than older children, and age differences in performance would become more pronounced on later trials.

To test whether the requirement to remember two rules or associations alone was sufficient to cause the younger children difficulty, we constructed a control version of our Stroop-like day–night test. Here, each card contained one of the two abstract designs. Children were instructed to say "day" to one design and "night" to the other. We predicted that subjects would perform significantly better on this control version than on the experimental version of the day–night task because the control version requires only memory, without also requiring inhibition. Specifically, we predicted that the percentage of correct responses would be higher, and response latency shorter, on the control version than on the experimental version, especially for younger children.

There have been a few attempts in the past to study performance on a task like the Stroop test in children. In particular, a study conducted by Passler, Isaac, and Hynd (1985) with children 6–12 years of age included a "verbal conflict" task in which subjects were asked to point to a gray card when the experimenter said "day" and to point to a white card when the experimenter said "night". Passler and colleagues found no significant differences in performance on this task over the 6–12 age range because children appeared to be performing at ceiling already by age 6. Our task shares some characteristics with Passler et al.'s task, but differences also exist. Passler and his colleagues used plain cards of a single solid color, white or gray. The cards we used (white with a picture of the sun and black with a picture of the moon and stars) are probably more strongly associated with day and night, and so our task should require more inhibition. Also, in Passler et al.'s task, subjects were only required to *recognize* the correct word for a certain card (the experimenter said "day" or "night" and the subject had to recognize which card belonged with that word), whereas our task requires subjects to *recall* the correct word for the card (the subject has to recall what word goes with each card without any reminder of the response possibilities). We expected to find a significant number of children having difficulty with our task, allowing us then to explore the sources of their difficulty, because we were studying younger children ($3\frac{1}{2}$–7 rather than 6–12 years) and because our task is more difficult.

2. Method

2.1. Subjects

We tested 240 normal, healthy children. One hundred and sixty children were tested on our Stroop-like day–night task with white-sun and black-moon cards (hereafter called the "sun–moon" or "experimental" condition). All were full-term and were from middle to upper-middle class families. Testing took place in a quiet area in the subjects' schools or in our laboratory at the University of Pennsylvania. Twenty children (10 male, 10 female) were tested at each age: $3\frac{1}{2}$, 4, $4\frac{1}{2}$, 5, $5\frac{1}{2}$, 6, $6\frac{1}{2}$, and 7 years. We used rather strict criteria for assigning children to a given age group: for example, only children between 3 years, 4 months of age and 3 years, 9 months were included within the $3\frac{1}{2}$-year age group. Only children between 3 years, 10 months and 4 years, 3 months were included in the 4-year age group. Similar cutoffs were used for all the other ages. The mean for each age group in weeks and days is provided in Table 1.

Initially, we tried to test 3-year-olds (mean age = 3 years [1 month]; range = 3[0]–3[4]) but we dropped this age group because the task appeared to be too difficult for them. Most of the 20 3-year-old subjects we tried to test either wouldn't play or failed the pretest. Besides the 160 children $3\frac{1}{2}$ years or older included in our analyses of the sun-moon condition, we attempted to test 10 children at $3\frac{1}{2}$ years, 10 children at 4 years, and 3 children at 5 years but they had to be excluded from the analyses because of experimental error (8 subjects), they wouldn't play (8 subjects), or they failed the pretest (19 subjects; see Table 2A).

We also tested 80 children on the control version of our Stroop-like day–night test using abstract designs (hereafter referred to simply as the control condition). The backgrounds of these children were the same as those for children tested in the sun-moon condition (see Table 3). All were full-term, healthy, and from middle to upper-middle class families. Testing for the control version also took place in a quiet area in the subjects' schools or in our laboratory. Twenty children (10 male, 10 female) were tested at each age: $3\frac{1}{2}$, 4, $4\frac{1}{2}$, and 5 years of age. The same criteria used for assigning children to age groups for the experimental condition were used here. (See Table 1 for the mean ages in weeks and days.) We did not test children on

Table 1
Mean age in weeks and (days) for each age group in each condition

	Age in years							
	$3\frac{1}{2}$	4	$4\frac{1}{2}$	5	$5\frac{1}{2}$	6	$6\frac{1}{2}$	7
Experimental condition	187(3)	212(2)	236(3)	262(4)	288(3)	316(3)	342(3)	363(2)
Control condition	185(4)	211(3)	235(4)	261(3)				

Table 2
Number of subjects who were excluded from analyses by age, sex, reason for exclusion, and condition

Age in years		Reasons why subjects were not used			
		Experimenter error	Child would not play	Child failed the pretest	Total number of subjects who could not be used
A: *Experimental condition*					
3	Male	2	3	3	8
	Female	1	2	4	7
$3\frac{1}{2}$	Male	1	1	4	6
	Female	1	1	2	4
4	Male	1	0	5	6
	Female	2	1	1	4
$4\frac{1}{2}$	Male	0	0	0	–
	Female	0	0	0	–
5	Male	0	3	0	3
	Female	0	0	0	–
B: *Control condition*					
$3\frac{1}{2}$	Male	0	1	1	2
	Female	0	1	1	2
4	Male	0	1	0	1
	Female	0	1	1	2
$4\frac{1}{2}$	Male	0	0	1	1
	Female	0	0	0	–
5	Male	1	0	0	1
	Female	0	0	0	–

Note: There were no unusable subjects in the $5\frac{1}{2}$–7-year age range.

Table 3
Demographic information on the subjects

	Experimental condition	Control condition
Mean birth weight (lb) (oz)	7(11)	7(11)
Mean number of siblings	1.6	1.8
Percentage of subjects with no siblings	10.7	4.0
Mean age of mothers at child's birth (years)	30.6	30.1
Mean age of fathers at child's birth (years)	33.0	32.5
Mean number of years of education of mothers	14.7	15.7
Mean number of years of education of fathers	15.4	16.2
Percentage of mothers working since child's birth	54.0	50.0

the control condition after 5 years of age because children were already performing at ceiling by age 5. To counterbalance card and instructions, 5 males and 5 females were tested at each age group with one card as "day" and an equal number were tested with the other card as "day". In addition to the 80 children included in our analyses, we tried to test 4 other children at $3\frac{1}{2}$ years, 3 children at 4 years, 1 child at $4\frac{1}{2}$ years, and 1 child at 5 years, but these children are omitted from the analyses below because of experimental error ($N = 1$), the children wouldn't play ($N = 4$), or they failed the pretest ($N = 4$; see Table 2B).

2.2. Materials

Two sets of cards were used – one set for the sun–moon condition and one set for the control condition. The dimensions of each card were 13.5×10 cm. There were 2 training cards and 16 testing cards in each set. The front of half of the cards for the experimental condition was black with a moon and stars. The front of the other experimental cards was white with a bright yellow sun (see Fig. 1a and b). The front of half of the cards for the control condition had a red and blue checkerboard pattern. The front of the other control cards had a blue background with two red squiggles that formed an X (see Fig. 1c and d).

2.3. Procedure

Training and pretest

The experimenter showed the subject a black moon card (or one of the control cards) and instructed the subject, "When you see this card, I want you to say 'day'." The experimenter asked the subject to repeat the word "day". The experimenter then removed the card and showed the white sun card (or the other control card), and instructed the subject, "When you see this card, I want you to say 'night'." The experimenter asked the subject to repeat the word "night".

The experimenter then showed the subject a white sun card (or one of the control cards). This time no instruction was given. If the subject hesitated, the experimenter prompted the subject by saying, "What do you say for this one?" The experimenter never said the words "night" or "day" as a prompt. If the subject responded correctly, the experimenter praised the child and proceeded to a practical trial with the black moon card (or the other card). If the subject responded correctly to the black moon card, the experimenter praised the child and these first two trials were then counted as trials 1 and 2 of testing; testing continued from there. If the subject responded incorrectly or did not respond at all on either of these trials, these two trials were counted as practice and the experimenter immediately reminded the subjected of both rules beginning with the card that the child had identified incorrectly. Then the experimenter started over, beginning

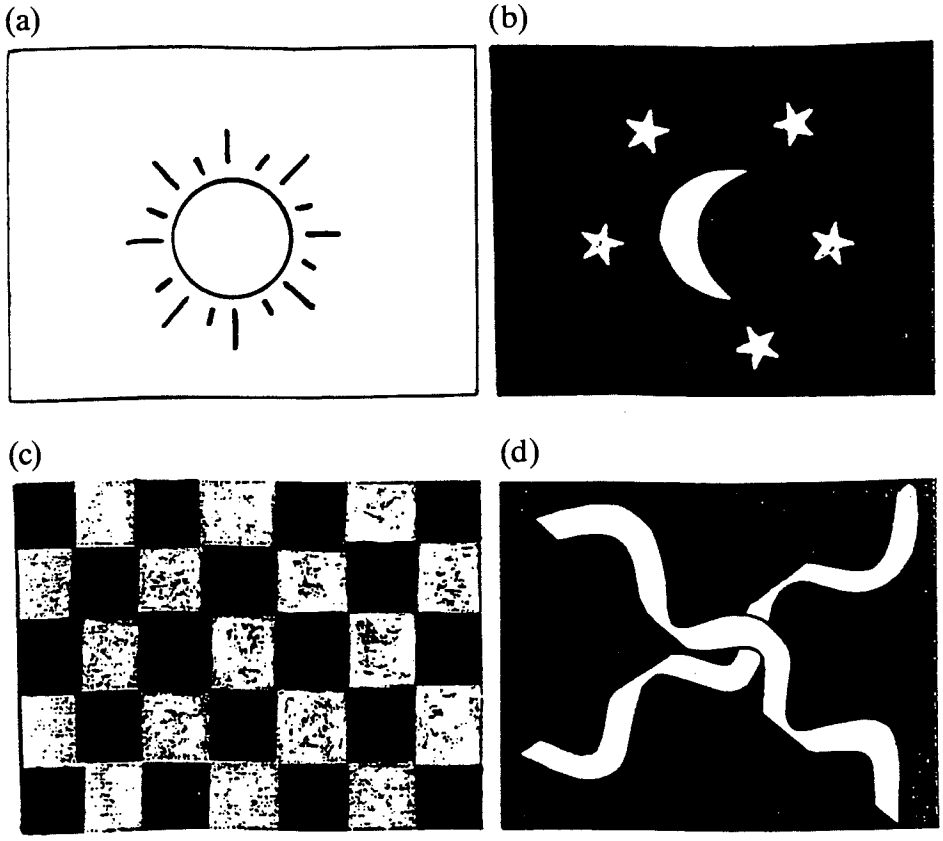

Fig. 1. Task stimuli. (a, b) Cards used for our Stroop-like day–night test for children. (a) Subjects were instructed to say "night" when shown this card. (b) Subjects were instructed to say "day" when shown this card. (c, d) Cards used for the control condition. (c) Half the subjects were instructed to say "night" when shown this card. Half were instructed to say "day". (d) Half the subjects were instructed to say "day" when shown this card. Half were instructed to say "night."

with the white sun card (or the corresponding control card). If the subject responded correctly, the experimenter praised the child and proceeded to a trial with the black moon card (or the other control card). If the subject was correct on this black moon card trial, these two trials were counted as trials 1 and 2 of testing and the experimenter continued from there. If the subject was wrong on either of these two trials, these trials were counted as practice and the experimenter immediately reminded the subject of the two rules beginning with the card that the child had identified incorrectly. Testing started from here. A subject needed to have answered each rule correctly at least once over the course of practice plus trials 1 and 2 in order for the session to be counted as usable. That is, we needed to see some evidence early on that the child understood what we were asking him or her to do in order for the session to count. The reason we counted early practice trials, if answered correctly, as part of testing, was that children who readily grasped

what we were asking became very bored if we tried to give them much practice.

Testing

Sixteen trials were administered in which eight "day" cards and eight "night" cards were presented according to a pseudorandom sequence. The cards were presented in the order night (n), day (d), d, n, d, n, n, d, d, n, d, n, n, d, n, d for both the experimental and control conditions. If the subject hesitated, the experimenter prompted the subject by saying, "What do you say for this one?" The experimenter never said the words "night" or "day" as a prompt. During the 16 test trials, no feedback was given to subjects.

In the control condition, half the subjects were told to say "day" when shown the squiggle card and to say "night" when shown the checkerboard card. The other half of the subjects were told to say "night" for the squiggle card and "day" for the checkerboard. We used both sets of instructions to insure that good performance on the control version was not due to any tendency to associate the words "day" or "night" with either of the abstract designs. We predicted that performance in the two versions of the control condition would be comparable.

We analyzed the results for each version of our Stroop-like task in a linear regression model to look for differences over age, sex, or condition (experimental vs. control; control 1 vs. control 2) and to determine if performance varied over trials within the same child. The dependent variables were whether a response on a given trial was correct or not, the number of correct responses over a session, response latency on each trial, and response latency over all trials within a session. Response latency was measured from the time the child first saw the "day" or "night" card until he or she gave a verbal response. It was coded from the videotape and intercoder reliability was .90.

3. Results

3.1. Performance in the sun–moon condition

A regression of percent correct on age, sex, and age × sex revealed that performance improved significantly with age, $F(1, 152) = 30.83$, $p < .01$; see Fig. 2. There was no significance difference between the sexes, $F(1, 152) = 0.02$, NS, and no significant interaction between age and sex, $F(1, 152) = 0.01$, NS). We also compared performance of the boys and girls at each age and found no significant difference in percentage correct at any age.

Response latency decreased significantly with age (linear regression: $F(1, 150) = 23.71$, $p < .01$). This age-related improvement in speed to respond in the sun–moon condition occurred primarily between $3\frac{1}{2}$ and $4\frac{1}{2}$ years of age (see Fig. 3). A regression model of reduced response latency

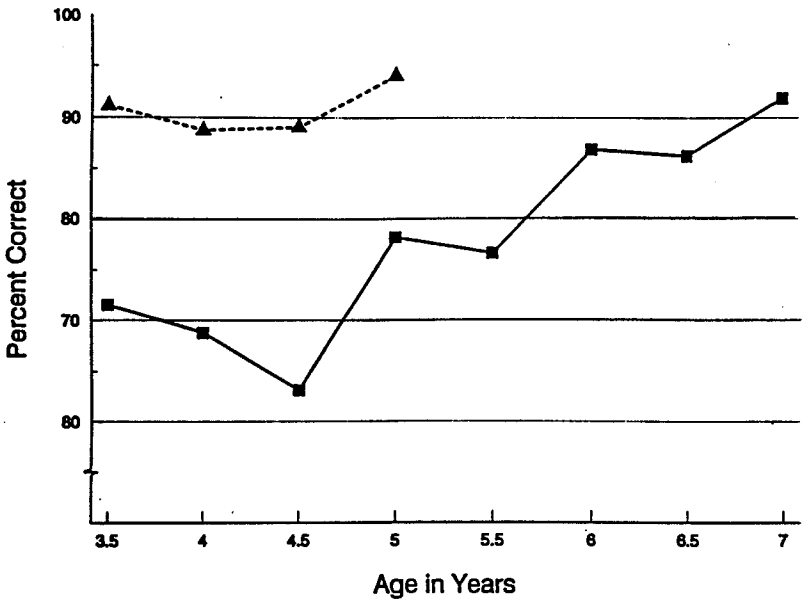

Fig. 2. Percentage of correct responses over age for the experimental and control conditions. $N = 20$ at each age in each condition. Number of trials = 16 for each session in each condition. Performance in the sun–moon experimental condition is indicated by the solid line; performance in the control condition by the dashed line. No subject older than 5 years was tested in the control condition.

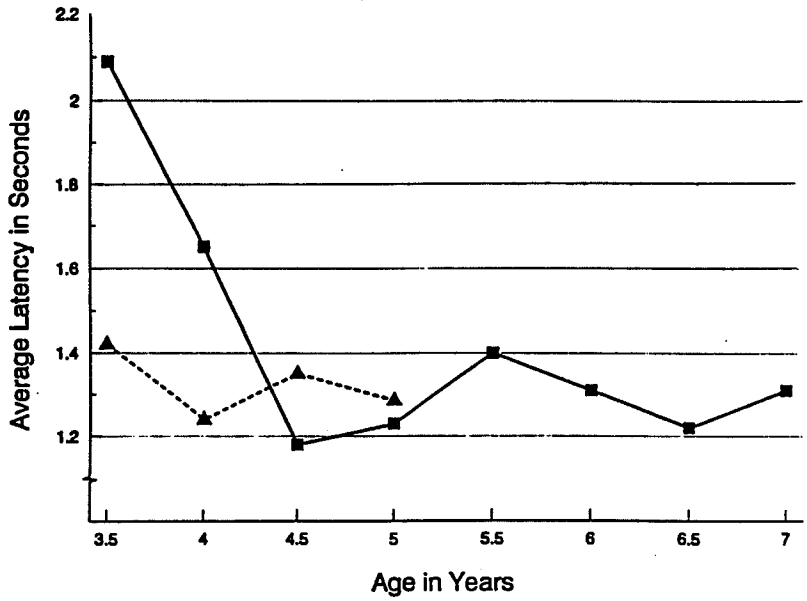

Fig. 3. Average time taken to respond by age and condition. Performance in the sun–moon condition is indicated by the solid line; performance in the control condition by the dashed line. No subject older than 5 years was tested in the control condition.

from $3\frac{1}{2}$ to $4\frac{1}{2}$ years and then no change thereafter (slope of zero from $4\frac{1}{2}$ to 7 years) fit the data significantly better than a regression model of a continuous, linear reduction in response latency from $3\frac{1}{2}$ to 7 years (test of the two models: $F(2, 3) = 25.00$, $p < .01$). There was a significant difference in the response latency of boys and girls, $F(1, 150) = 5.76$, $p < .05$; boys were faster. This sex difference was only significant, however, at one individual age ($3\frac{1}{2}$ years).

Children at all ages performed well at the outset of testing. Overall percentage correct for the first four trials was 87.5% (see Table 4). As predicted, even the younger children ($3\frac{1}{2}$–$4\frac{1}{2}$ years) performed well here (mean percentage correct on trials 1–4 for younger subjects = 81.0%; for older subjects = 91.5%).[1] Also as predicted, age differences in the performance became more pronounced on later trials (mean percent correct on the last 4 trials for younger subjects = 54.3%; for older subjects = 78.3%). The difference between younger and older children in the percentage of correct responses on the last four trials (24%) was double that on the first 4 trials (12%).[2] Performance of *all* children deteriorated over the course of a session, however. Across all ages, children gave more incorrect responses on later trials than on earlier ones (regression of response accuracy on trial number (1–16): $F(1, 2204) = 44.35$, $p < .01$). Similarly, percentage correct on the first four trials was significantly higher than percentage correct on the last four trials over all ages (paired $t(158) = 7.45$, $p < .01$), and at each individual age. The deterioration in performance over the course of a session was more pronounced for the younger children, however. This can be seen by (1) the significant interaction of age × trial number in the regression of percentage correct on these variables, $F(1, 2204) = 8.02$, $p < .01$, (2) the significant effect of age in the regression of the difference (percentage correct on first four trials minus percentage correct on last four trials), $F(1, 158) = 5.67$, $p < .05$; see Fig. 4, and (3) the significant orthogonal contrast comparing $3\frac{1}{2}$–$4\frac{1}{2}$ year-olds versus 5–7-year-olds on this difference, $F(1, 158) = 6.92$, $p < .01$. That is, the difference in percentage of correct responses on the first four trials versus the last four trials was significantly greater for younger children than for older children.

Over the course of a session, children also began responding more quickly (regression of response time on trial number: $F(1, 1978) = 16.44$, $p < .01$).

[1] Our decision to define "younger" as $\leq 4\frac{1}{2}$ years and "older" as ≥ 5 years was based on the finding that response latency decreased markedly from $3\frac{1}{2}$ to $4\frac{1}{2}$ years, remaining stable thereafter. There was no marked discontinuity at any point in the developmental progression in percentage of correct responses. We have also analyzed the data with a midpoint split ($3\frac{1}{2}$–5 vs. $5\frac{1}{2}$–7) and obtained similar results in all analyses.

[2] Most subjects could sustain successful performance for more than two trials but performance tended to deteriorate after four trials, hence percentage correct on the first four trials provides a reasonable indication of good performance early in a session. To compare performance early and late in a session we chose an equal number of trials at the session's end. In each of these sets of four trials, each rule was tested twice.

Table 4
Percentage of correct responses and mean time to respond across age and condition over the whole session (16 trials) and over the first and last four trials

Age in years	Mean percentage correct	Mean percentage correct on first four trials	Mean percentage correct on last four trials	Mean response latency	Mean response latency on first four trials	Mean response latency on last four trials
$3\frac{1}{2}$	71.5	83.8	60.0	2.09	2.61	1.91
	91.1	*98.8*	*88.8*	*1.42*	*1.76*	*1.25*
4	68.8	85.5	56.6	1.65	2.09	1.42
	88.8	*94.0*	*81.7*	*1.24*	*1.46*	*1.22*
$4\frac{1}{2}$	63.1	73.8	46.3	1.18	1.42	0.96
	89.1	*93.8*	*87.5*	*1.35*	*1.45*	*1.32*
5	78.1	87.5	73.8	1.23	1.41	1.19
	94.1	*98.8*	*95.0*	*1.29*	*1.33*	*1.23*
$5\frac{1}{2}$	76.2	87.5	73.5	1.40	1.41	1.19
6	86.9	92.5	76.25	1.31	1.38	1.32
$6\frac{1}{2}$	86.3	95.0	82.9	1.22	1.34	1.25
7	91.9	96.3	87.5	1.31	1.29	1.34

Experimental Stroop condition = values in regular type (top line).
Control condition = values in italics (second line).
$N = 20$ at each age in each condition.

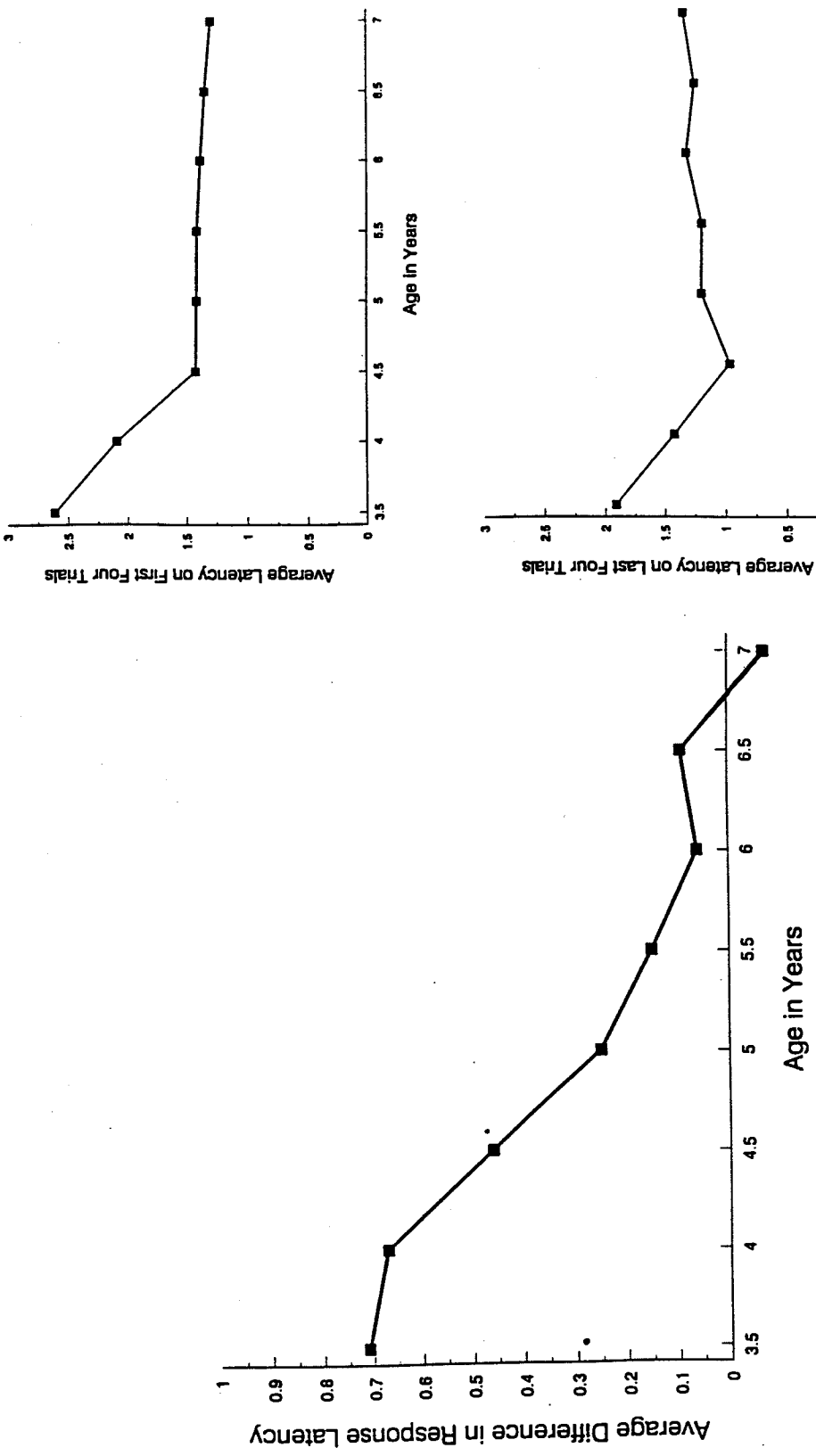

Fig. 4. Difference in accuracy early and late in the session in the experimental condition of our Stroop-like day–night task for children: percentage of correct responses on the first four trials minus percent correct on the last four trials by age and condition. The difference for each subject was calculated and then the average was calculated for each age.

The same can be seen in the comparison of response latency on the first and last four trials: response latency was significantly longer on the first four trials than on the last four trials of that same session (paired $t(152) = 4.32$, $p < .01$; see Table 4). This decrease in response time over the course of a session was especially true for younger children. Indeed, the response latency of children $4\frac{1}{2}$ years of age and older did not change significantly over the course of a session, and the difference in response latency (trials 1–4 vs. trials 13–16) was no longer significant after 5 years of age (see Fig. 5).

Children performed better on the trials where they took longer to respond, $r(2135) = .044$, $p < .05$. The relation between accuracy and response latency was particularly pronounced for younger children. For ages $3\frac{1}{2}$–$4\frac{1}{2}$ years, the mean percent correct on the first four trials was 80.0% and the mean response latency for these trials was 2000 ms. The mean percent correct on the last four trials for this age group was 54.3% and the mean response latency for these trials was 1400 ms. When younger children took longer to answer, they were more often correct. For children aged 5–7 years, mean percent correct on the first four trials was 91.5% and mean response latency for these trials was 1400 ms. Their mean percentage correct on the last four trials was still high (78.3%) and their mean response latency was roughly the same as earlier in the session (1300 ms).

The number of trials needed to pass the pretest decreased with age (linear regression: $F(1, 154) = 31.63$, $p < .01$). The number of unusable subjects also decreased significantly with age ($F(1, 8) = 20.19$, $p < .01$; see Table 2A). Both of these findings further suggest that our task was easier for the older children.

When we added the following demographic variables (birth weight, mother's education, father's education, mother's occupation, father's occupation, socio-economic status, whether and how long a child was in daycare, birth order, and number of siblings) to our regression equations, we found no significant main effects or interactions with a single exception: Children who had been in day care were correct on significantly more trials than children who had been cared for at home, $F(1, 96) = 9.91$, $p < .01$).

3.2. Performance on the control condition

As predicted, there was no significant difference between the two control conditions in percentage of correct responses, $t(78) = 1.42$, $p > .1$, or in response latency, $t(78) = .48$, $p > .1$. Indeed, at each age, performance was comparable in the two conditions on both of these measures. We have, therefore, combined the data from both control conditions for all analyses reported below.

Percentage correct did not differ significantly by age (linear regression: $F(1, 79) = 0.96$, NS; see Fig. 2, nor did response latency ($F(1, 78) = 1.35$, NS; see Fig. 3). There was little difference in performance over age because,

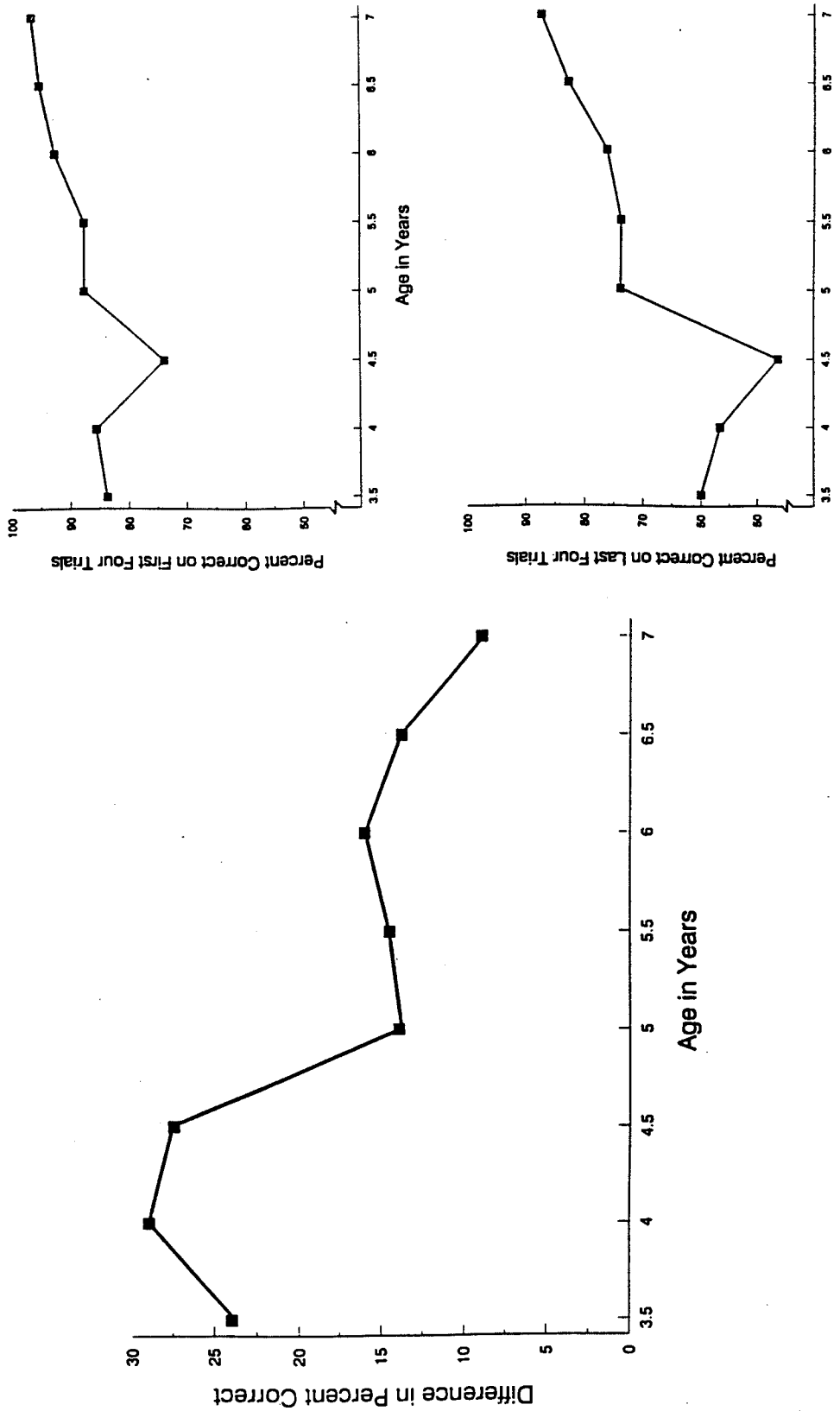

Fig. 5. Difference in speed of responding early and late in the session in the experimental condition of our Stroop-like day–night task for children: mean response latency on the first four trials minus latency on the last four trials. The difference for each subject was calculated and then the average was calculated for each age.

as predicted, the control task was easy for even the youngest subjects; at each age percentage correct was greater than 85%.

There was a significant difference between boys and girls in percentage correct (girls performed better; $F(1, 79) = 4.80$, $p = .03$), but at no individual age was the difference between the performance of boys and girls significant. We found no sex difference in response latency, $F(1, 78) = 0.78$, NS.

The number of trials needed to pass the pretest for the control condition decreased with age, $F(1, 79) = 4.43$, $p = .03$.

Performance of the younger children deteriorated over the course of a session even on this easier task. Overall, children gave more correct responses on earlier trials than on later ones (regression of percent correct on trial number: $F(1, 1251) = 12.60$, $p < .01$). This worsening of performance over a session was no longer significant after $4\frac{1}{2}$ years of age. Analysis of performance on the first four trials versus the last four trials yielded similar results (paired $t(78) = 3.47$, $p < .01$), with analyses at each individual age showing that the difference in performance at the beginning and end of the session was significant only at the two youngest ages ($3\frac{1}{2}$ and 4 years).

Children also responded more quickly as a session progressed. This can be seen in the significant decrease in response latency over the course of a session (regression of response latency on trial number: $F(1, 1251) = 33.29$, $p < .01$) and in the significant difference between response latency on the first four versus the last four trials (paired $t(77) = 4.19$, $p < .01$). Analyses at each individual age revealed that the regression of response latency on trial number was no longer significant after $4\frac{1}{2}$ years of age and the comparison of latency on the first and last four trials no longer yielded a significant difference after 4 years of age.

3.3. Comparison of performance in the control and experimental conditions

As predicted, children performed significantly better in the control condition (as assessed by their percentage of correct responses) and found the control condition significantly easier than the experimental condition (as assessed by their response latency and their ease in passing the pretest). Percentage correct for the control condition was significant higher than for the experimental condition ($t(158) = 7.31$, $p < .01$; see Fig. 2). This was true at every age ($3\frac{1}{2}$ years: $t(38) = 3.09$, $p < .01$; 4 years: $t(38) = 4.22$, $p < .01$; $4\frac{1}{2}$ years: $t(38) = 4.63$, $p < .01$; 5 years: $t(38) = 3.07$, $p < .01$) and for males and females ($t(78) = 4.17$, $p < .01$; $t(78) = 6.52$, $p < .01$, respectively).[3] Response latency was significantly shorter for the control version than for

[3] To compare the control and experimental conditions of the task, we used only subjects between the ages of $3\frac{1}{2}$ and 5 years because children outside this age range were not tested on the control condition.

the experimental version ($t(152) = 2.72$, $p < .01$; see Fig. 3). We interpret this as another indication that the control condition was easier for our subjects since we take longer response latency to indicate that subjects had to work harder in order to formulate their response. This difference in response time by version was significant, however, only at the two youngest ages ($3\frac{1}{2}$ and 4 years) ($t(38) = 3.56$, $p < .01$; $t(35) = 2.86$, $p < .01$; respectively).

More children failed the training for the experimental condition (12 subjects) than for the control condition (4 subjects) (binomial distribution 12 vs. 4, $p < .03$). Also, subjects needed more practice trials before they were able to pass the pretest for the experimental condition than for the control condition, $t(158) = 2.54$, $p < .05$.

Performance fell off more sharply in the experimental condition over the course of a session than in the control condition. The difference between percentage correct earlier in a session (the first four trials) and performance later in that same session (the last four trials) for sun–moon condition was significantly larger than that same difference for the control condition (t-test of the difference in percentage correct for the experimental condition vs. the control condition: $t(158) = 3.32$, $p < .01$). When we analyzed this difference in percentage correct at each age, we found a significant difference only at $4\frac{1}{2}$ years, $t(33) = 2.16$, $p < .05$. This difference in percentage correct was significant for girls, $t(77) = 3.37$, $p < .01$, but not for boys, $t(76) = 1.27$, NS.

The difference in response latency early versus late in a session for the experimental condition was also significantly larger than the same difference for the control condition, $t(149) = 2.33$, $p < .05$, but this difference between the conditions was not significant at any individual age. It was significant for girls, $t(75) = 2.61$, $p < .05$, but not for boys, $t(71) = 1.02$, NS. Response latency changed more over the course of a session among children in the experimental condition because they started out taking so very long to respond. Children never took this long in the control condition (latency on the first four trials, experimental vs. control conditions: $t(151) = 3.15$, $p < .01$), and their response time remained more constant over the course of a session.

3.4. Types of errors and trends

One common error was for children to alternate in the their responses during the course of a session. In other words, subjects began to seemingly mindlessly alternate in saying "day" to one card, "night" to the next card, then "day" to the next, etc. These subjects did not wait to clearly see the card before giving their response, despite the experimenter's repeated urging that they wait to see the card before responding. More younger subjects tended to alternate than older subjects (regression: $F(1, 159) = 8.41$, $p < .01$). On average, 4.3 subjects between the ages of $3\frac{1}{2}$ and $4\frac{1}{2}$ alternated, whereas only an average of 1.6 subjects 5–7 years old alternated.

Indeed, only 2 subjects between the ages of $5\frac{1}{2}$ and 7 alternated, while 18 subjects below $5\frac{1}{2}$ years did so. All who alternated were tested on the experimental condition. Across all ages, those subjects who fell into this pattern of alternating tended to do so after correct responses of "day", "night", "day", "night" on the previous four trials, specifically after trials 7 and 13.

Another common error was for subjects to "match" their response to the card. For example, they said "day" to the white sun card and "night" to the black moon card. The tendency to say what a card really represented rather than its opposite decreased significantly with age (regression: $F(1, 8) = 32.12$, $p < .01$). Nine subjects between the ages of $3\frac{1}{2}$ and 5 consistently matched on at least nine of the 16 trials, two $5\frac{1}{2}$-year-old subjects did so, and no subjects over the age of $5\frac{1}{2}$ years did so. Almost all errors consisted of giving the "matching" errors since subjects rarely answered with a word other than "day" or "night" and rarely gave no response at all.

Another error that some children made was to say either "day" or "night" to every card throughout the session. Other subjects said various other words (i.e., "morning time," "sun," "moon") to one of the cards while saying "day" or "night" to the other card.

4. Discussion

Our predictions were confirmed. Younger children had considerable difficulty with the day–night task when the black moon and white sun cards were used. This can be seen in their relatively low percentage of correct responses, their long response latencies, and their difficulty in passing the pretest. The mean percentage of correct responses at all ages <6 years was less than 80%, whereas children of 6–7 years were correct on approximately 90% of the trials. The youngest children ($3\frac{1}{2}$ and 4 years of age) also showed extremely long response latencies (approximately 2000 ms); older children took only approximately 1000 ms to respond. Several children of $3\frac{1}{2}$–4 failed the pretest; no child over $4\frac{1}{2}$ years did so. Older children passed with relative ease; younger children who managed finally to pass the pretest needed more practice trials on average to do so than older children. If we had included all younger children in our analyses, even those who failed the pretest, we would probably have found even more pronounced age differences in performance than are reported here. The age-related increase in percentage of correct responses was relatively continuous over the $3\frac{1}{2}$–7 age range, but the decrease in speed of responding occurred primarily by $4\frac{1}{2}$ years.

The younger children who passed the pretest, started out performing well on the initial trials, but were not capable of sustaining a high level of performance over the subsequent trials. Perhaps their performance declined

because the experimental task was so difficult for them and consumed so much of their cognitive energy that they could not maintain such a high level of effort throughout the 16 trials of testing. One indication of the great effort the task demanded of younger children is the long time it took them to respond. Sometimes people take longer to respond when they don't know an answer. In these cases, response latency tends to be longer on trials where subjects perform more poorly. We noticed, however, that on this task younger children tended to do *better* on trials where they took longer. When younger children took the time they needed, they succeeded on the task. The longer latencies shown by the younger subjects might indicate that they were working harder (i.e., needed to churn their "mental gears" longer) than older subjects. Perhaps younger children needed to put forth so much effort to perform their task correctly that they could not sustain a high level of performance because they exhaust themselves on the first few trials. One might almost conceive of this as if all the children started out with an equal allotment of "energy". Because the task was difficult for the younger children they exhausted their allotment early in the session; whereas older children needed less effort on each trial and so had plenty of "energy" left by the end of the session.

The difference in the latencies of younger and older children is due primarily to the difference early in the session. Toward the end of a session, children of all ages were responding quickly (in roughly 1300 ms), but the younger children were no longer responding correctly. Older children, on the other hand, did not have to work as hard (their response latency was short even early in a session) and their percentage of correct responses remained high throughout. Here we are hypothesizing: (1) the long response latency for the younger subjects early in a session indicates that they were working very hard (and during this time their performance was good); (2) the short response latency for younger subjects later in a session is indicative of their having given up (and their percentage correct fell to near chance); (3) we take the short response latency of the older subjects throughout the session to indicate that the task was relatively easy for them (and their percentage correct remained high throughout). Similarly, we take the relatively short response latency of even the younger subjects in the control condition to indicate that this condition was relatively easy for them (and their percentage correct throughout the session was significantly higher than in the experimental condition).

Let us turn now to heart of the matter: what accounts for the poor performance of the younger children in the experimental condition of the Stroop-like day–night task? It is not that young children did not understand the instructions, for the children included in our analyses passed the pretest and even the youngest children performed fairly well on the four initial test trials. Neither was the problem for the younger children an inability to remember two rules or associations, for the control condition also required memory of two rules and even the youngest children performed the control

task well. A short attention span cannot account for the difference in performance between the experimental and control tasks either because both tasks had the same general instructions, same number of trials (16), and indeed same everything except for the stimulus cards, yet younger children could sustain a high level of performance in the control condition but not in the experimental one. The reason is, of course, that the experimental task was more difficult. Why was it more difficult? Why might it have required more attention?

One possibility is that the moon–sun day–night task requires children to exercise inhibitory control over their behavior. It is possible that inhibiting the tendency to say what the picture represents is so difficult for young children that they cannot keep this up after the initial trials. One reason younger children had such difficulty passing the pretest may have been their difficulty inhibiting the matching response even for a moment.[4] The long time that younger subjects needed to formulate their response when they were correct may also indicate that they needed to expend considerable effort to inhibit the more natural response. In addition, almost all errors consisted of giving the "matching" response, which might indicate that subjects could not inhibit saying "day" to the sun and "night" to the moon. However, since there were only two response possibilities[5] there was little opportunity to err except by "matching," and since subjects rarely erred on most or all trials it is difficult to know for sure whether the errors were random or systematic. Thus, we cannot be certain that their intermittent errors were due to lack of inhibition rather than to inattention or forgetfulness.

Certainly there is evidence that 3- and 4-year-old children have difficulty exercising inhibitory control over their behavior. For example, in the "windows task" of Russell, Mauthner, Sharpe, and Tidswell (1991), children were rewarded when they pointed to a box which they could see was empty, and were *not* rewarded when they pointed to a box in which they could see candy. Children of 3 years were unable to inhibit the tendency to point to the baited box. In another study (Zelazo, Frye, & Reznick, submitted), children were asked to sort a deck of cards by one criterion and then by another. Children of 3 years did well on the first sorting criterion but had difficulty switching, despite the experimenter's instructions that the sorting rule had changed and what the current rule was, *and despite the child's demonstrated understanding and memory of that rule.* For example, when asked where red things should go, children of 3 years pointed to the

[4] We required that children pass the pretest because we felt it important that they demonstrate some understanding of what we were asking them to do. However, children could fail the pretest *either* because they did not understand what they were to do *or* because of an inability to *demonstrate* this understanding due to their inability to inhibit the customary response.

[5] Subjects rarely answered with a word other than "day" or "night" and rarely gave no response at all.

correct place. Immediately thereafter, however, the experimenter handed a red card to the child, saying, "Where does this red thing go?"; the child sorted it by the previously correct criterion, shape. That is, children of 3 years seemed unable to inhibit sorting by the criterion that had been correct. Similarly, Livesey and Little (1971) and Bell and Livesey (1985) found that children 3 and 4 years of age were unable to inhibit inappropriate responses and so performed incorrectly, but this was not due to lack of knowledge as the children could verbalize the correct answer. See also Kopp (1982).

Indeed, Diamond (e.g., Diamond & Gilbert, 1989; Diamond, 1990, 1991b) has argued that to a surprising extent even infants can figure out, and remember, what they are supposed to do, but their inability to inhibit more automatic reactions gets in the way of their demonstrating what they know. That is, to some extent the problem is not "cognitive", in the sense of it being inadequate reasoning or memory. Rather, the problem is in gaining control over one's behavior, in going from cognition to action:

> Thus, infants and frontal patients sometimes show an apparent dissociation between what they know and what they demonstrate in their behavior; their behavior appears to be captured by more automatic, prepotent response tendencies that are not inhibited as they should be. Avoiding such errors requires keeping your intention firmly in mind, and controlling your behavior so that it expresses what you intend. . . . [I]nfants appear to know more than their behavior indicates. As their ability to exercise inhibitory control increases, cognitive abilities are revealed that may have been present for some time. (Diamond, 1990:1ii–1iii)

Two variations of our task could be tried to determine if reducing the inhibitory requirement of the task makes it substantially easier for younger children. One variation would be to use cards that are just black and white, or just gray and beige, without any suns, moons, or stars. Presumably these solid color cards would not be as strongly associated with day and night as are the cards in our present task. Another possible modification of the task would be to instruct subjects to say less standard words than "day" and "night," such as "morning" and "evening". Presumably these words might not be as strongly associated with the white sun card and the black moon card.

Another possible reason why younger children may have found our task so difficult is that the task requires *both* memory *and* inhibition. Perhaps children are able to remember two rules and to continue to inhibit their natural response, but they are unable to sustain good performance when the two requirements are combined. Many of the tasks that children of 3–5 years fail may be thought of as requiring subjects to keep two things in mind plus inhibit a natural or predominant response. An appearance–reality task (e.g., Flavell, 1986) might require, for example, keeping in mind that a sponge made to look like a rock is both a sponge and a rock, and that although it looks like a rock it is really a sponge. (The inclination that must be inhibited is to give the response that matches one's perception – the

inclination to say that if it looks like a rock it is a rock.) One may perhaps think of theory of mind problems (e.g., Wimmer & Perner, 1983) in a similar way. For example, Mary must keep in mind both where the hidden object is now and where Billy saw it placed before, and she must inhibit her inclination to say where the object really is and instead say where Billy would think it is, even though she shows that this answer is "wrong" because the object is not there now.

Perhaps younger children experienced a cognitive overload, not because of the exact combination of memory plus inhibition, but because needing to do anything in addition to remembering two rules was too much for them. This idea could be explored, for example, through a variation of our task that requires memory of three rules. If cognitive overload were the problem, younger subjects should perform poorly on this task just as they performed poorly in our experimental condition. However, if inhibition, or inhibition plus memory, is the problem then younger subjects might still perform well, even when asked to remember three rules.

One might conceive of "overload" in terms of exceeding available mental "disk" space (e.g., younger children might have only enough mental bins to hold two items but older children might have three or four bins; see, e.g., Pascual-Leone, 1970; Case, 1972). Or, one might imagine that the memory of younger children fades faster or is fuzzier than that of older children (as in a poorer signal:noise ratio), so that younger children have to work harder to hold onto what they are trying to keep in mind. A faint or fuzzy memory of the two associations might suffice if there are no additional demands, but the need to remember a third association or inhibit a strong response tendency might overtax the system.

A final explanation for the developmental improvement we found might be that older children simplified the task and remembered only one rule (i.e., to say the opposite) while younger children tried to remember two rules (i.e., to say "day" to the black moon card and to say "night" to the white sun card). Some subjects, especially older ones, said to the experimenter that the game was to "say the opposite". This idea could be explored in our task by explicitly instructing subjects to "say the opposite" to each card. If remembering two rules is the problem, younger subjects should perform better with this new procedure than they did with the procedure used here. If inhibition is the problem, younger children should perform at the same level with this new procedure as they did with the present procedure.

One way to avoid children simplifying the task by encoding only to "say the opposite" is to require them to say a word unrelated to the pictures. For example, the experimenter might instruct the child to say "dog" to the black moon card and "horse" to the white sun card. Here, each required response is unrelated to the picture and unrelated to the other response. Inhibition would still be required on this task in that the subject would have to inhibit the tendency to say what the picture on the card represents (unlike our

control condition in which the pictures were abstract and therefore not associated with any particular label). It is not clear, however, whether inhibiting saying "day" in order to say "horse" is easier or harder than inhibiting saying "day" in order to say "night".

We interpret the longer response latency at the younger ages, and the correlation between longer response latency and correct responses at the younger ages, to mean that children under 5 years needed more time to perform well on our task than did older children. Certainly, there is much evidence that speed of processing appears to increase as children get older (e.g., Kail, 1988, 1991a, 1991b; Rose, Gottfried, Melloy-Carminar, & Bridger, 1982). However, did the children need more time on this Stroop-like task to formulate their response (after seeing the card for that trial) or did they need more time to settle down before the next trial? If they needed more time to formulate their response, imposing a mandatory delay after each card is shown and before the child responds on each trial should improve performance. If they needed more time between trials to clear their minds and prepare for the next trial, placing a mandatory delay between trials should improve performance.

What can be conclude from the particular errors that younger children made? Some subjects said either "day" or "night" to every card. These children may not have remembered either rule. They might have remembered that they were supposed to say "day" to one of the cards, for example, but forgotten to which one. On the other hand, the children may have remembered one rule correctly, but forgotten the other. For example, if subjects said "day" throughout a whole session, they might have remembered that they were supposed to say "day" to the black moon card, but forgotten that they should say "night" to the white sun card. Some subjects consistently said the correct word to one card and consistently said a different word or gave no response to the other card. For example, one girl of $3\frac{1}{2}$ years tested on the control stroop task had a problem during practice remembering the "day" rule, but finally was able to pass the pretest. She remembered both rules during the first half of the session, but then forgot the checkerboard = day rule. Throughout the second half of the session, whenever the checkerboard card was shown, she told the experimenter that she forgot what to say to that card. She was correct on *all* of the squiggle = night cards throughout the entire session but failed to say "day" to any of the checkerboard cards after trial 7. Another example is one subject who consistently said "day" to the black moon card and consistently said "morning time" to the white sun card. Such subjects appear to have forgotten one rule, but appear to have correctly remembered the other.

There is another possible explanation for why some children said various other words (e.g., "morning time") to one of the cards. Usually these subjects said a synonym for one of the correct responses. For example, they said "morning time" when the correct answer was "day", or "dark" when

the correct answer was "night". Perhaps they were using the "say the opposite" rule discussed above. Subjects may have said synonyms to the correct answer because they remembered the general rule but had forgotten the exact terminology.

Our results agree with those of Passler et al. (1985) in that we, too, found little change in performance after 6 years of age. By the age of 6, subjects were already performing near ceiling. However, by testing subjects younger than those tested by Passler et al., we were able to demonstrate that before 6 years the abilities required by our task undergo marked development changes. Whereas children of $3\frac{1}{2}$–$4\frac{1}{2}$ years have a terribly difficult time with the task, by 6–7 years the task is trivially easy. Older children were able to sustain good performance throughout the course of a session while younger children could not. We suggest that this was because older children did not have to work as hard as the younger ones. Perhaps the task was easier for the older children because they simplified the two rules we presented to one rule and so had to remember less, or perhaps because they did not need to work as hard to inhibit the "matching" response. Since the results show that children of all ages had little trouble with the control condition, we conclude that memory of two rules alone is not the problem for younger children. Our experimental version of this Stroop-like task requires memory of two rules plus inhibition of a prepotent response. Hence, the reason younger children found the task difficult was probably because (a) inhibiting a prepotent response is difficult for them, (b) doing anything in addition to remembering two things is difficult for them, or (c) the specific conjunction of inhibiting a prepotent response and a significant memory load is beyond their ability.

What we have tried to do in this paper is describe a task that we developed and that is easy to administer to children, chart the developmental progression in children's performance of the task, and begin to explore why the task is difficult for younger children. We have explored how performance changes over trials on the task, how the performance by trial function changes over age, and how the relationship between response latency and accuracy changes over trials and over age. We have been able to rule out two possible interpretations for why younger children do not perform better (that they didn't understand the instructions (ruled out by their passing the pretest and performing well at the outset of testing) or that they cannot remember two rules or associations (ruled out by their excellent performance in the control condition)). When we changed but one aspect of the task (the stimulus cards), the task became trivially easy even for the youngest subjects. We think the fact that the task requires children to keep two things in mind and to stop themselves from saying what the cards really represent accounts for why young children find the task so difficult. More work is needed, though, before we will know whether our interpretation is the correct one.

References

Bell, J.A., & Livesey, P.J. (1985). Cue significance and response regulation in 3- to 6-year old children's learning of multiple choice discrimination tasks. *Developmental Psychobiology, 18,* 229–245.

Case, R. (1972). Validation of a neo-Piagetian capacity construct. *Journal of Experimental Child Psychology, 14,* 287–302.

Diamond, A. (1988). The abilities and neural mechanisms underlying AB̄ performance. *Child Development, 59,* 523–527.

Diamond, A. (Ed.), (1990). *The development and neural bases of higher cognitive functions. Annals of the New York Academy of Sciences, 608.*

Diamond, A. (1991a). Frontal lobe involvement in cognitive changes during the first year of life. In K.R. Gibson & A.C. Petersen (Eds.), *Brain maturation and cognitive development: comparitive and cross-cultural perspectives* (pp. 127–180). New York: Aldine de Gruyter.

Diamond, A. (1991b). Neuropsychological insights into the meaning of object concept development. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: essays on biology and knowledge* (*pp.* 67–110). Hillsdale, NJ: Erlbaum.

Diamond, A., & Gilbert, J. (1989). Development as progressive inhibitory control of action: retrieval of a contiguous object. *Cognitive Development, 4,* 223–249.

Diamond, A., & Goldman-Rakic, P.S. (1986). Comparative development in human infants and infant rhesus monkeys of cognitive functions that depend on prefrontal cortex. *Society for Neuroscience Abstracts, 12,* 742.

Diamond, A., & Goldman-Rakic, P.S. (1989). Comparison of human infants and rhesus monkeys on Piaget's AB̄ task: evidence for dependence on dorsolateral prefrontal cortex. *Experimental Brain Research, 74,* 24–40.

Diamond, A., Zola-Morgan, S., & Squire L. (1989). Successful performance by monkeys with lesions of the hippocampal formation on AB̄ and object retrieval, two tasks that mark developmental changes in human infants, *Behavioral Neuroscience, 103,* 526–537.

Flavell, J.H. (1986). The development of children's knowledge about the appearance–reality distinction. *American Psychologist, 41,* 418–425.

Kail, R. (1988). Developmental functions for speeds of cognitive processes. *Journal of Experimental Child Psychology, 45,* 339–364.

Kail, R. (1991a). Developmental change in speed of processing during childhood and adolescence. *Psychological Bulletin, 109,* 490–501.

Kail, R. (1991b). Processing time declines exponentially during childhood and adolescence. *Developmental Psychology, 27,* 259–266.

Kopp, C.B. (1982). Antecedents of self-regulation: a developmental perspective. *Developmental Psychology, 18,* 199–214.

Livesey, D.J., & Little, A. (1971). Sequential learning by children. *Journal of Genetic Psychology, 118,* 33–38.

Luria, A.R. (1973). *Higher cortical functions in man.* New York: Basic Books.

Pascual-Leone, J.A. (1970). A mathematical model for transition in Piaget's developmental stages. *Acta Psychologia, 32,* 301–345.

Passler, P.A., Isaac, W., & Hynd, G.W. (1985). Neuropsychological development of behavior attributed to frontal lobe functioning in children. *Developmental Neuropsychology, 4,* 349–370.

Perret, E. (1974). The left frontal lobe of man and the suppression of habitual responses in verbal categorical behavior. *Neuropsychologia, 12,* 323–330.

Piaget, J. (1954). *The construction of reality in the child.* New York: Basic Books. Original French edition, 1937.

Rose, S.A., Gottfried, A.W., Melloy-Carminar, P., & Bridger, W.H. (1982). Familiarity and novelty preferences in infant recognition memory: implications for information procession. *Developmental Psychology, 18,* 704–713.