A Longitudinal Study of Young Children's Memory Behavior and Performance in a Sort-Recall Task

WOLFGANG SCHNEIDER

Max Planck Institute for Psychological Research

AND

BEATE SODIAN

University of Munich

One hundred and sixty-nine children were tested twice, at the ages of about 4 and 6 years, for their recall of a list of clusterable objects, following either play-and-remember or sort-and-remember instructions, and were assessed for their metamemory of the efficacy of conceptual and perceptual sorting strategies. In addition, an independent sample of 30 4-year-olds was tested twice within a time interval of 2 weeks to assess short-term stability of the memory measures. The main developmental trends reported in cross-sectional studies of preschoolers' memory were replicated. Conceptual clustering during recall significantly predicted recall performance in 6-year-olds, regardless of instructional condition, whereas in 4-year-olds verbal IQ and conceptual clustering predicted recall only under sort-and-remember instructional conditions, memory span being the only significant predictor under play-and-remember instructions. Although the short-term stabilities of all variables except for metamemory were high, over the 2-year period under study the stabilities of the memory variables were low with the exception of free recall. An analysis of individual stabilities showed that the more children recalled at age 4 the more stable their recall was over the 2-year period. These results indicate that although memory behavior and performance may be considerably less stable in the preschool than in the elementary school years, differential patterns may emerge at an early age. © 1991 Academic Press, Inc.

Preschool age thus seems to be a formative period for the emergence of strategic competencies which fully develop during the elementary school years. While 4-year-olds seem to rely mostly on simple perception-bound strategies such as naming or close visual examination of the items to-be-remembered (Baker-Ward, Ornstein, & Holden, 1984), 6-year-olds

14

employ conceptual organization at encoding and retrieval to a higher degree than 4-year-olds who do so only when explicitly instructed to sort (Sodian, Schneider, & Perlmutter, 1986). Furthermore, not only the use of memory strategies but also their effectiveness seems to increase over the preschool years (Baker-Ward et al., 1984). However, this interpretation of memory development in young children is based exclusively on cross-sectional findings. While some first (short-term) longitudinal evidence is available on memory development in older children (Kunzinger, 1985), no such studies have been reported in preschool and kindergarten age. The lack of longitudinal data limits the inferences that can be drawn from existing research in important ways (see Ornstein, Baker-Ward, & Naus, 1988; Schneider & Weinert, 1989). In particular, cross-sectional studies do not yield information on the stability of interindividual differences in intraindividual change. The increase in clustering scores and recall performance observed in cross-sectional studies between the ages of 4 and 6 years could, for instance, be due to the fact that most children in a sample have made some progress in organizational behaviors and memory performance or, that some children have made enormous progress while most children have remained the same or even declined.

Another shortcoming of cross-sectional studies is that they do not allow the prediction of interindividual differences despite general developmental change. Thus, we do not know whether a 4-year-old's level of performance on a memory task relative to that of other 4-year-olds allows us to predict that child's relative position in his or her age group at age 6 or later. Longitudinal studies of grade-school children's memory development provided some first information on these issues. Kunzinger (1985), for instance, found that rehearsal set size at the first measurement point was not related to recall at this measurement point but predicted recall 2 years later. A particularly high level of individual stability for rehearsal set size indicated that those children who initially showed larger set sizes maintained their position relative to the group 2 years later. No corresponding evidence has been reported for preschool children.

The present study attempts to gather some first information on the stability and predictability of selected aspects of metamemory, memory behavior, and memory performance in 4- to 6-year-old children. Children were tested for free recall, conceptual clustering at encoding and retrieval, and task-related metamemory in a sort-recall task similar to the one used by Sodian et al. (1986) at the age of 4 years, and again approximately 2 years later. As studies on memory development in this age range are rare and have generally been conducted with small samples, one aim of the present investigation was to replicate the findings from the cross-sectional study by Sodian et al. (1986) in a larger sample.

The second aim of the longitudinal investigation was to study the

stability of children's recall, clustering, and metamemory over a 2-year-period. Cognitive performance in preschool children has generally been assumed to be more "fragile" than in older children, being affected by minor situational variations. Thus, young children's performance in memory tasks may be unstable even over a short period of time. Despite these assumptions, short-term stability of cognitive performance has hardly ever been assessed in studies of young children's memory. Such an assessment is necessary in a longitudinal study to determine whether memory behavior and performance and metamnemonic knowledge can be measured reliably. If short-term stability of the relevant measures is high, low long-term stability of memory behaviors and performance can be interpreted as a genuine instability of performance over the period under study. If, however, short-term stability is low, long-term instability may reflect measurement problems rather than genuine instability. We therefore tested an independent sample of 4-year-old children twice within 2 weeks with the same instruments used in the longitudinal study.

Group stability of memory behaviors and performance refers to stability at the aggregate level and is typically measured by the correlation coefficient. As aggregate stability does not allow inferences on the individual level (Asendorpf, 1989; Valsiner, 1986), it is necessary to additionally estimate individual stabilities and their variance in a population (i.e., differential stability). Individual stability refers to constancy of relative position (cf. Wohlwill, 1973, p. 361ff.), that is, to the constancy of an individual's standing relative to some referent group across age. Such a measure is the lability score (Bayley, 1949), which is the across-age standard deviation of an individual's z scores.

The present study is a first attempt to investigate both group and individual stabilities of preschoolers' recall, clustering, and metamemory over a 2-year-period. We expected group stabilities to be lower than those observed in grade-school children. The aim of the analysis of individual stabilities was to gather some information on the possible reasons for low stability of memory performance in preschoolers: Is most preschoolers' performance in memory tasks unstable over a period of 2 years or are some children extremely unstable while the majority shows high stability? If a subgroup of extremely unstable children can be identified, how can this subgroup be characterized?

As our study is part of a larger longitudinal investigation, we were able to address this issue by relating memory performance in the sort-recall task under study to a number of other theoretically interesting variables: Children's text recall was assessed in two different stories (Knopf & Waldmann, 1987). We also obtained measures of memory capacity as well as measures of verbal and nonverbal intelligence. The Columbia Mental Maturity Scale (Burgemeister, Blum, & Lorge, 1972) was used to assess nonverbal intelligence, and the verbal scale of the

Hannover-Wechsler Intelligence Scale for Preschool and Kindergarten Children (Eggert, 1978) was chosen as a measure of verbal intelligence. Further, the word span task developed by Case, Kurland, and Goldberg (1982) was used as an indicator of memory capacity.

METHOD

Subjects

The subjects were 169 children (81 boys and 88 girls) who participated in the Munich Longitudinal Study on the Genesis of Individual Competencies (cf. Weinert & Schneider, 1987). From the 173 subjects available at the beginning of the study, four children dropped out during the first year. The children were tested for the first time when they were approximately 4-years-old (mean age = 4;3; range 3;6 to 4;7), and for the second time 2 years later when they were about 6-years-old (mean age = 6;3; range 5;6 to 6;7). In addition, 30 4-year-olds, 15 boys and 15 girls (mean age = 4;2; range 3;5 to 4;5) were tested twice within a time interval of 2 weeks.

Materials and Procedure

The stimuli were 16 small, brightly colored toys, approximately uniform in size. The list contained one red, yellow, green, and blue item from each of four categories: animals (blue elephant, green dog, yellow cow, red horse), furniture (blue wardrobe, green chair, yellow table, red bed), vehicles (blue bus, green tractor, yellow truck, red car), and household items (blue teapot, green frying pan, yellow cup, and red pot).[1]

Subjects were tested individually in a quiet room in their kindergarten. Half the children were randomly assigned to a "play-and-remember" condition and the other half to a "sort-and-remember" condition. At the second measuring each child was administered the task under the same instructional conditions as at the first measuring point.

Memory Task

The procedure was identical to that used by Sodian et al. (1986). After all the toys were presented, children in the *Play-and-remember* condition were told that they were allowed to play with the toys for a short while. Thereafter, the experimenter would hide the toys in the box, and children had to tell the experimenter the names of the toys.

---

[1] The stimulus list was adopted from Sodian et al. (1986). Sodian et al. (1986) used two parallel lists and did not find any list effects. Due to various organizational problems with our large-scale longitudinal study, the same list of items was given to all subjects of the longitudinal sample. This was not true for the subjects of the "short-term stability" study who were presented with two sets of materials which were counterbalanced from Time 1 to Time 2. Again, no list effects were found.

Children in the *Sort-and-remember* condition were instructed that they should put all those toys together that go together. They were then told that the experimenter would hide the toys in the box later on, and that they would have to remember the names of the toys.

In the play/sort phase, children were allowed to play with the objects or to sort them for 2 min. After 2 min a color photograph that showed the final arrangement of the objects was taken. Then the toys were hidden and children were asked to tell the names of all the toys they had seen.

Metamemory task. The experimenter told the child that she had also played the game in which the child had just participated with other children, and that she wanted to show the child what the other children had done to remember the toys. She then presented three color photographs showing (from left to right) a random organization of the items, an organization by semantic category, and an organization by color. She then pointed out to the subject that the photographs showed what different children had done to remember the toys, and explained that one child "had just somehow grouped the toys together" (random organization), that another child "had grouped them so that all the animals, all the vehicles, all the furniture, and all the household things were together" (organization by semantic category), and yet another child "had grouped them so that all the blue things, all the yellow things, all the red things, and all the green things were together" (organization by color). The experimenter then asked the children to show her how they can remember the toys best. When the children had picked a photograph the experimenter took this photograph and asked the children to show her how they can remember the toys second best. Then children were asked to select the photograph showing which was the worst way to remember the toys. Finally, the experimenter asked the children to justify their choice of the best way to remember.

RESULTS

Preliminary analyses of variance on the recall and organization data involving age as a within-subject factor and instructional condition and sex as between-subject factors revealed no systematic effects of sex. Thus the data were collapsed across this variable.

Free Recall

Table 1 shows the mean number of items correctly recalled in the play-and-sort instruction conditions separately for each measurement point. An analysis of variance on these data was conducted involving age as a within-subject factor and instructional condition as a between-subject factor. The analysis revealed a significant main effect of age, $F(1, 167) = 365.55$, $p < .01$. There was no effect of instructional condition, and no significant interaction. Not surprisingly, children recalled

TABLE 1

MEAN NUMBER OF ITEMS CORRECTLY RECALLED AND MEAN CONCEPTUAL CLUSTERING SCORES (ENCODING AND RECALL) IN THE PLAY-AND-REMEMBER AND SORT-AND-REMEMBER INSTRUCTIONAL CONDITIONS, SEPARATELY FOR EACH MEASUREMENT POINT

| | Instructional condition | | | | | |
| | Play-and-remember | | | Sort-and-remember | | |
Age	Correct recall	RR in encoding	RR in recall	Correct recall	RR in encoding	RR in recall
4	6.89	.22[a]	.42	7.34	.39	.44
	(2.57)	(.19)	(.23)	(2.76)	(.28)	(.20)
6	10.96	.28	.44	11.60	.54	.54
	(2.15)	(.26)	(.19)	(2.48)	(.25)	(.17)

Note. Standard deviations are in parentheses.

[a] Clustering scores of .39 and above are significantly greater than chance level.

more items at the second measurement point than they recalled at measurement point 1 (11.28 vs. 7.11). Although the mean number of items correctly recalled in the sort condition was more than in the play condition (9.47 vs. 8.32), and approached significance in the 4-year-olds ($p < .10$), this difference did not prove to be significant.

Organization during Encoding and Recall

In order to assess the degree of color and category organization during encoding in the two instructional conditions, the photographs showing the final arrangements of the objects (after the 2-min play or sort period) were analyzed using a procedure first developed by Kee and Bell (1981). The number of categorical repetitions (i.e., either horizontal or vertical juxtapositions of exemplars from the same taxonomic group or color) was counted for each photograph. A ratio of repetition (RR) measure (see Bousfield & Bousfield, 1966) was computed for both category and color clustering. The chance value derived from the formula for the expected number of repetition was .2 for our list of 16 items, with a maximum clustering score of .8 (cf. Murphy, 1979). The RR measure was chosen because it has been shown statistically to be relatively independent of the level of total recall and of the size and numbers of categories recalled (cf. Murphy & Puff, 1982).

Color clustering scores for both encoding and recall were small and at chance level, regardless of instructional condition and measurement point. Thus these data will not be considered further. Mean RR conceptual clustering scores for encoding and recall in the two instructional conditions are also shown in Table 1, separately for each measurement point. A 2(instructional condition) × 2(age) repeated measurement analysis of variance on conceptual clustering during encoding yielded main

effects of instructional condition, $F(1, 167) = 19.07$, $p < .01$ and age, $F(1, 167) = 16.01$, $p < .01$. The age × instructional condition interaction did not reach statistical significance. Children showed significantly more conceptual clustering during encoding at measurement point 2 than they did 2 years earlier (.41 vs. .31), and RR scores above chance level were only obtained in the sort-and-remember instructional condition (.47 vs. .25).

As to conceptual clustering during recall, a 2(instructional conditions) × 2(age) repeated measurement analysis of variance revealed significant main effects of instructional condition, $F(1, 167) = 7.25$, $p < .01$ and age, $F(1, 167) = 10.57$, $p < .01$. However, the main effects were qualified by a significant age × instructional condition interaction, $F(1, 167) = 4.23$, $p < .05$. Simple-effects tests based on the overall error term from the analysis of variance revealed that the children in the sort-and-remember instruction showed more conceptual clustering than the subjects in the play-and-remember condition at the second measurement point, $t(167) = 3.69$, $p < .01$, whereas the two instructional groups did not differ from each other at the first measurement point.

Metamemory

At the first measurement point 48 children ranked conceptual organization highest, while 93 ranked organization by color and 28 random organization highest. The corresponding numbers for the second measurement point were 51, 114, and 4, respectively. No differences were observed between instructional conditions at either measurement point. The majority of the children ranked color clustering highest at both measurement points. No significant age trends were observed. Of those children who preferred conceptual clustering, 71% at the first measuring point and 98% at the second point were able to correctly justify their choice. However, short-term stability of metamemory judgments was extremely low for 4-year-olds: 21 out of 30 children who were assessed twice with the same instruments as the subjects of the longitudinal study changed their preference over a 2-week-interval. Thus, the possibility to interpret the metamemory data as indicating stable preferences appears to be limited. Consequently, these data will not be considered in all further analyses (see Table 2 for short-term stabilities of the other measures).

Stability over Time

Group stability. Of major interest to the present study is the extent to which the various cognitive variables remain stable over time.

Group stabilities (test-retest correlations) for the various memory variables used in this study are given in Table 2.

For the purpose of comparison, group stabilities for the two intelligence

TABLE 2

TEST-RETEST CORRELATIONS (GROUP STABILITIES) FOR THE MEMORY VARIABLES AND IQ
MEASURES USED IN THE STUDY, SEPARATELY FOR THE SORT-AND-REMEMBER AND
PLAY-AND-REMEMBER SUBJECTS

	Sort-and-remember	Play-and-remember
Free recall	.36[a]	.37[a]
	(.68)	(.71)
Conceptual clustering	.07	.17
(encoding)	(.85)	(.78)
Conceptual clustering	.22[a]	.01
(recall)	(.54)	(.68)
Preference for conceptual clustering	.11	.00
(metamemory)	(.24)	(.04)
Memory span	.29[a]	.21[a]
Verbal IQ (HAWIVA)	.60[a]	.56[a]
Nonverbal IQ (CMMS)	.57[a]	.55[a]

Note. Short-term stability coefficients are in parentheses.
[a] Correlations significant at the .05 level.

measures (i.e., verbal and nonverbal intelligence) are also shown in Table 2. It is apparent from this table that most memory variables were not stable between 4 and 6 years of age. With the exception of free recall, stability coefficients were very low for both instructional conditions, thus indicating that most individuals considerably changed their relative standing within their group between the two measurement points.

In order to decide whether the amount of instability found for the memory variables was due to unreliability of measures or to true fluctuation/change in the variables, their short-term stability was additionally assessed. That is, an independent sample of 30 4-year-old children was given the sort-recall task twice within a time interval of about 2 weeks. The resulting coefficients of short-term stability are given in parentheses in Table 2. They show that most memory variables used in this study could be reliably assessed. For these variables, long-term instability in the various memory measures seems to be due to differential rates of individual changes between 4 and 6 years of age. The only exception was the metamemory measure which showed insufficient stability, regardless of experimental condition.

Individual stability. To test the assumption of differential rates of individual changes further, *individual* stability was also assessed. As the difference between group stability and individual stability may not be immediately apparent, it will be briefly discussed below. It is crucial to distinguish between the correlation, in a sample of persons, of two assessments of the same memory variable at different points in time and the correlation, in a sample of variables, of two assessments of the same

person at different points in time. As noted by Ozer and Gjerde (1989), the former approach is variable-centered and describes the degree of consistent ordering of individuals over time; the latter approach is referred to as person-centered and describes the degree to which the ordering of attributes within each individual remains constant over time relative to the other attributes. The latter approach has also been termed *constancy of relative position* (Wohlwill, 1973) and is referred to as *individual stability* in this paper.

According to Wohlwill (1973), the most general representation of individual stability is the amount of across-age variability shown in an individual's relative standing within the referent group. A traditional measure is the lability score (Bayley, 1949), which is the across-age standard deviation of an individual's z scores. To cope with one undesirable property of this score, that is, its skewed distribution, we used the coefficient of individual stability developed by Asendorpf (1989) which is based on a strictly monotonic transformation of the individual's z scores normalizing the skewed distribution of the individual stabilities. One specific advantage of this transformed score is that correlating the transformed individual stabilities with external variables of interest no longer poses a problem. Correlating individual stabilities with external variables seems important because it may facilitate the explanation of differential stability and also gives information about possible regression-towards-the-mean effects in the data.

Because of space limitations, we focus on the individual stabilities of the free recall variable, which was the only memory measure showing sufficient group stability in both experimental conditions. For the sake of clarity, the untransformed individual stability coefficients (z scores) are given to illustrate the amount of lability over time observed in the recall data. While small z-score differences obtained for an individual's recall data at the first and second measurement point indicate considerable constancy of the individual's standing relative to the reference group, large z-score differences reflect individual instability over time. The mean lability score (0.89) found for the two recall assessments indicates that the absolute z-score differences comprised almost a standard deviation, thus reflecting considerable individual instability over time. Differential stability, as indicated by the standard deviation of the mean individual stability coefficient, is also substantial in the sample (0.68). In order to explain the differential stabilities, their dependence on absolute recall scores as well as on verbal and nonverbal intelligence was assessed via correlation. There was only one significant correlation of $r = .24$, $p < .05$, between the transformed individual stabilities and free recall at the first measurement point. This finding indicates that the more children recalled at the age of 4, the more stable their recall proved to be over the 2-year period.
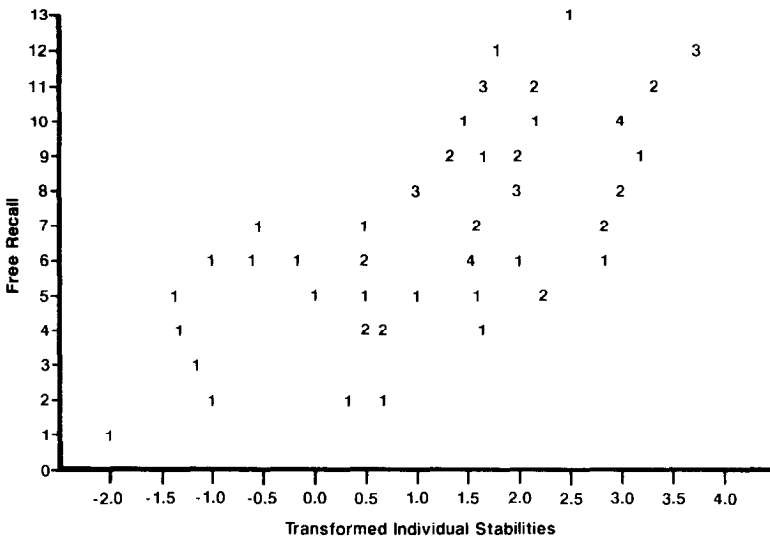
Note. Numbers displayed on scatterplot indicate number of observations at each position

FIG. 1.   Scatterplot of the bivariate distribution of children's individual stabilities in free recall and their initial scores in this variable.

For a closer analysis, the scatterplot of the correlation of r = .24 was investigated (cf. Fig. 1). The scatterplot shows that individual stabilities tended to be particularly low for those children scoring low at the first measurement point. On the other hand, most children scoring above the sample mean at the first measurement point demonstrated relatively great stability over time.

As noted by Asendorpf (in press), correlations are very sensitive to extreme scores in the bivariate distribution. That is, a few extreme but unstable scores can obscure a substantial aggregate stability in the rest of the sample (and vice versa). A closer look at Fig. 1 reveals that this also holds true for our data: Obviously, a minority of children with extreme scores at one occasion contributed considerably to the substantial aggregate instability of scores.

Subgroup Comparisons

In our subgroup analysis, we focused on the empirical finding that high instability was only observed for those subjects showing low recall at Time 1. Inspection of the data revealed that most of these subjects at least doubled their recall from Time 1 to Time 2. Not surprisingly, these children's recall scores tended to be in the normal range at Time 2.

While this finding could be due to true developmental change in children's mnemonic skills, an alternative explanation is that we were not

able to assess true competence at Time 1. If assessment of true memory competence is the crucial problem, this should also be reflected in unstable recall scores on similar memory tasks presented at the *same* measurement point. We were able to test this prediction, because two additional text recall measures were available for Time 1 that required the children to listen to a story and then to recall it as accurately as possible. Both text recall measures tapped similar contents ("birthday party" vs. "play in the afternoon") and were highly intercorrelated in the total sample ($r = .67$).

We compared two subgroups, formed on the basis of the longitudinal recall data. The first subgroup consisted of 24 children who scored low in the sort-recall task at Time 1 and at least doubled their performance from Time 1 to Time 2. This group was compared to a subgroup of 18 children with high recall at Time 1, that is, initial scores at least one standard deviation above the group mean. As expected, these two subgroups differed considerably regarding long-term stability of recall in the sort-recall task ($r = .67$ vs. .29 for the high- and low-scoring groups, respectively). Even more interesting, text recall obtained for the initially low-scoring subgroup (hereafter referred to as Low Recall/Low Stability group) differed from that obtained for the initially high-scoring children (High Recall/High Stability group) in a way that points to measurement error problems. Intercorrelations between the two text recall measures were .51 and .86 for the Low Recall/Low Stability and High Recall/High Stability groups, respectively. About 71% of the Low Recall/Low Stability subjects were very inconsistent on the two text recall measures, with scores for one text recall at least double those for the other text recall, whereas only 7% of the High Recall/High Stability subjects were classified as inconsistent based on this criterion. It appears, then, that the extreme variability in the recall data of the Low Recall/Low Stability children at Time 1, and the instability over time observed for this subgroup should be interpreted more as measurement problems in assessing true competence than as fluctuation in rates of true developmental change.

Table 3 illustrates the major differences and commonalities between the two subgroups. As can be seen from Table 3, High Recall/High Stability subjects showed superior recall for most memory measures. It seems important to note, however, that the two subgroups did not differ regarding conceptual clustering, which was similarly low for both groups. Interestingly enough, verbal and nonverbal IQs were roughly comparable and in the normal range for the two subgroups. Thus it appears that problems with assessing true mnemonic skills in the Low Recall/Low Stability subjects are not related to their general intellectual abilities. As we suspected, the small group of unstable children was responsible for the low overall group stability observed for the recall data. Omitting the

TABLE 3
MEAN RECALL, MEMORY SPAN, AND IQ AS A FUNCTION OF SUBGROUP (LOW RECALL/LOW
STABILITY VERSUS HIGH RECALL/HIGH STABILITY)

	Group			
	Low recall/	High recall/		
Variables	low stability	high stability	t	p
Free recall (Time 1)	4.27 (1.55)	11.60 (1.50)	12.50	<.01
Free recall (Time 2)	11.27 (2.07)	13.50 (1.43)	3.52	<.01
Text recall 2	11.62 (7.60)	14.50 (13.46)	0.76	ns
Text recall 1	6.42 (6.63)	12.50 (9.35)	2.09	<.05
Memory span	2.00 (1.68)	3.11 (0.33)	2.70	<.05
Conceptual clustering (Time 1)	0.22 (0.17)	0.34 (0.31)	1.94	ns
Conceptual clustering (Time 2)	0.28 (0.26)	0.38 (0.24)	1.54	ns
Verbal IQ	105.71 (13.18)	108.25 (10.87)	0.56	ns
Nonverbal IQ	109.41 (10.53)	109.44 (11.57)	0.00	ns

Note. Standard deviations are given in parentheses.

10 most unstable subjects from the sample raised the group stability coefficient from .37 to .58. Excluding all the 24 low recall/low stability children from the sample further raised group stability ($r = .65$).

To explore the reasons for the difficulties with assessing the Low Recall/Low Stability subjects' true competencies, we carried the analyses one step further. A closer examination of the text recall data of this subgroup revealed that lower scores were always obtained for the first of two parallel assessments. As the sort-recall task was also given at the very beginning of the longitudinal study, one possible explanation for the poor performance of these children is that they felt particularly uncomfortable with interacting with unfamiliar experimenters in a new situation. This is supported by shyness assessments conducted by Asendorpf (1987): Of the 16 children in the Low Recall/Low Stability subgroup who participated in the shyness assessments (contact with an adult stranger), 13 were classified as very shy at the first measurement point. The hypothesis that the shy children were also the youngest subjects in the sample did not hold: only three of the shy subjects were under age 4 at the beginning of the study.

DISCUSSION

The purpose of this study was to replicate cross-sectional findings on memory development in preschool age in a larger sample, and to obtain information on the stability of individual differences in organizational

behaviors and memory performance. The results will be discussed with regard to these two issues.

Consistent with cross-sectional findings on young children's organizational behaviors in a sort-recall task (Sodian et al., 1986), even some 4-year-olds were shown to be able to conceptually sort items both at encoding and retrieval. While perceptual clustering scores were very low in both age groups, conceptual clustering scores were well above chance level under sort-and-remember instructional conditions even in 4-year-olds. Both the degree of conceptual organization and recall performance were shown to increase over the preschool years as predicted from cross-sectional findings.

The results of the present study emphasize the need to attend to individual differences in memory behavior and performance. That is, the comparison of the present findings with those obtained by Sodian et al. (1986) in a cross-sectional study with a small sample shows that caution is warranted in interpreting group differences and correlation coefficients based on small samples, as few children with extreme scores may considerably influence the means and correlation coefficients.

The present longitudinal study yielded some first information relevant to the following issues:

(1) How stable are individual differences in memory behavior and performance over the preschool years?

(2) Is instability of performance due to measurement error or to genuine fluctuation of performance?

(3) How can individual differences in *stability* be interpreted?

As expected, young children's stability scores were lower than those reported for elementary school children (Kunzinger, 1985). Our results for the preschool and kindergarten period indicate that individual differences in memory skills do not prove stable over time. As short-term stability scores were sufficient with the exception of metamemory, we can interpret these findings as reflecting largely genuine instability, not measurement error. It should be noted, however, that short-term stability found for our memory measures did not reach those of the intelligence measures. Thus, unreliability in the memory measures could to some extent have depressed long-term group stability scores.

The extreme unreliability of the metamemory measure in the short run points to a fundamental problem in the assessment of metamemory in preschoolers. Metamemory interviews in preschoolers almost always employ simple forced-choice formats. In the present task, these may have been interpreted by the children as referring to their personal preference for one or another pictorial arrangement of toys and not to the function of these arrangements for *remembering* the items. Thus, the metamemory

data do not seem to indicate a lag between children's use of organizational strategies and their conscious awareness of their effectiveness. Rather, young children's knowledge of the effectiveness of organizational strategies remains to be explored with more sensitive assessment procedures.

The analysis of *individual* stabilities provided some possible explanations for the low overall group stabilities of the memory measures. As illustrated for the case of free recall, instability seemed to be particularly high for subjects scoring low at the initial measurement point. Stability over time seems to be positively correlated with initial recall in the sense that most children already showing above average recall at the first measurement point tended to maintain their relative standing within the reference group. Additional subgroup analyses revealed that the relatively substantial stability of this group of children is obscured by the lability of a minority of the sample with extreme scores in the bivariate distribution.

Results for the Low Recall/Low Stability subgroup of 4-year-olds were difficult to interpret at first glance. Although these children were principally able to understand and follow the experimenter's instruction, they obviously did not always do so. Additional analyses including information on children's social anxiety provided an explanation for children's poor performance at the first measurement point: The majority of Low Recall/Low Stability subjects was classified as very shy and there is reason to believe that these children had difficulties with interacting with the unfamiliar experimenters. It appears, then, that the instability in these children's recall data over time cannot simply be reduced to a statistical regression-toward-the-mean effect, but has to be linked to specific personality characteristics which offer a substantive explanation for the phenomenon.

Note that it is very difficult to identify these children in cross-sectional research because parallel measures tapping the same construct are rarely used. In our case, the inclusion of children with initially very low recall scores in the longitudinal analysis considerably biased our estimate of true developmental change. Omitting these subjects from analysis leads to a different pattern of results. That is, although we were not able to replicate Kunzinger's (1985) finding that individual stability over time was *generally* high, our analysis revealed that it was high for the majority of children in our sample and particularly high for a *subgroup* of children, namely those subjects with initially higher recall scores.

Taken together, then, our results illustrate the importance of longitudinal studies of early memory development. As they allow for in-depth analyses of individual and differential stabilities of performance over time, they provide us with a comprehensive picture of interindividual differences in developmental change, thereby going beyond the information usually obtained from cross-sectional studies.

REFERENCES

Asendorpf, J. (1987). Social competence. In F. E. Weinert & W. Schneider (Eds.), *The Munich Longitudinal Study on the Genesis of Individual Competencies (LOGIC), Report No. 2: Documentation of assessment procedures used in waves one to three* (Technical Report). Munich: Max Planck Institute for Psychological Research.

Asendorpf, J. (1989). Individual, differential, and aggregate stability of social competence. In B. H. Schneider, G. Attili, J. Nadel, & R. Weissberg (Eds.), *Social competence in developmental perspective* (pp. 71–86). Dordrecht/NL: Kluwer.

Asendorpf, J. (in press). The measurement of individual consistency. *Methodika.*

Baker-Ward, L., Ornstein, P. A., & Holden, D. J. (1984). The expression of memorization in early childhood. *Journal of Experimental Child Psychology, 37,* 555–575.

Bayley, N. (1949). Consistency and variability in the growth of intelligence from birth to eighteen years. *Journal of Genetic Psychology, 75,* 165–196.

Bousfield, A. K., & Bousfield, W. A. (1966). Measurement of clustering and of sequential constancies in repeated free recall. *Psychological Reports, 19,* 935–942.

Burgemeister, B., Blum, L., & Lorge, J. (1972). *Columbia mental maturity scale.* New York: Harcourt Brace Jovanovich.

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology, 33,* 386–404.

Eggert, D. (1978). *Hannover-Wechsler-Intelligenztest für das Vorschulalter (HAWIVA).* Bern: Huber.

Kee, D. W., & Bell, T. S. (1981). The development of organizational strategies in the storage and retrieval of categorical items in free-recall learning. *Child Development, 52,* 1163–1171.

Knopf, M., & Waldmann, M. R. (1987). Memory for texts. In F. E. Weinert & W. Schneider (Eds.), *The Munich Longitudinal Study on the Genesis of Individual Competencies (LOGIC), Report No. 2: Documentation of assessment procedures used in waves one to three* (Technical Report). Munich: Max Planck Institute for Psychological Research.

Kunzinger, E. L. (1985). A short-term longitudinal study of memorial development during early grade school. *Developmental Psychology, 21,* 642–646.

Murphy, M. D. (1979). Measurement of category clustering in free recall. In C. R. Puff (Ed.), *Memory, organization, and structure* (pp. 51–83). New York: Academic Press.

Murphy, M. D., & Puff, C. R. (1982). Free recall: Basic methodology and analyses. In C. R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 99–128). New York: Academic Press.

Ornstein, P. A., Baker-Ward, L., & Naus, M. J. (1988). The development of mnemonic skill. In F. E. Weinert & M. Perlmutter (Eds.), *Memory development: Universal changes and individual differences* (pp. 31–50). Hillsdale, NJ: Erlbaum.

Ozer, D. J., & Gjerde, P. F. (1989). Patterns of personality consistency and change from childhood through adolescence. *Journal of Personality, 57,* 483–507.

Schneider, W., & Weinert, F. E. (1989). Memory development: Universal trends and individual differences in memory development. In A. de Ribaupierre (Ed.), *Transition mechanisms in child development: The longitudinal perspective* (pp. 68–106). Cambridge: Cambridge University Press.

Sodian, B., Schneider, W., & Perlmutter, M. (1986). Recall, clustering, and metamemory in young children. *Journal of Experimental Child Psychology, 41,* 395–410.

Valsiner, J. (1986). Between groups and individuals. Psychologists' and laypersons' interpretations of correlational findings. In J. Valsiner (Ed.), *The individual subject and scientific psychology* (pp. 113–151). New York: Plenum Press.

Weinert, F. E., & Schneider, W. (Eds.) (1987). *The Munich Longitudinal Study on the Genesis of Individual Competencies (LOGIC), Report No. 2: Documentation of as-*

sessment procedures used in waves one to three (Technical Report). Munich: Max Planck Institute for Psychological Research.

Wohlwill, J. F. (1973). *The study of behavioral development.* New York: Academic Press.