Journal of Autism and Developmental Disorders, Vol. 19, No. 4, 1989

The Role of Age at Assessment, Developmental Level, and Test in the Stability of Intelligence Scores in Young Autistic Children[1]

Catherine Lord[2]
Department of Pediatrics, University of Alberta and Glenrose Rehabilitation Hospital

Eric Schopler
Division TEACCH, Department of Psychiatry, University of North Carolina

Longitudinal comparisons were made of intelligence and developmental quotient (IQ/DQ) scores for three age groups of 70–72 autistic children aged 2 to 3, 4 to 5, and 6 to 7 years at initial assessment and reassessed at least 2 years later. Stability and predictability over a mean follow-up period of 5 years were related to age, developmental level, and test used at initial assessment. IQ/DQs during preschool years were quite stable and predictive of later IQ scores, except when early scores on the Bayley Scales of Mental Development were compared with later scores on performance or nonverbal tests. As for other populations, predictability for young autistic children was strongest when the same test was used at both assessments, and when children were 4 years or older at initial assessment.

The performance of autistic children on intelligence tests has been shown to be as stable and predictable from later preschool years to school age as that of nonautistic children with communication handicaps and/or behavior disorders (Freeman, Ritvo, Needleman, & Yokota, 1985; Lockyer & Rutter,

[1]This research was funded in part by grants from the Natural Sciences and Engineering Research Council of Canada and the Alberta Heritage Foundation for Medical Research to the first author. Preliminary results from this project were presented at the annual TEACCH Conference on Autism in Durham, North Carolina, in May 1986.
[2]Address all correspondence to Catherine Lord, Greensboro TEACCH Clinic, 1020 E. Wendover, Greensboro, NC 27405.

0162-3257/89/1200-0483$06.00/0

1969; Lord & Schopler, 1989). However, the intellectual assessment of children with autism, especially those who are very young (3 years and younger), those with no language or very limited communication skills, and those with severe delays in other areas is not a straightforward process (Rutter, 1985). In particular, little is known about the stability or predictability of the scores of children with estimated cognitive levels under 3 years, since most available tests would render these children "untestable" (Freeman et al., 1985; Rutter & Lockyer, 1967; Shah & Holmes, 1985).

It is our belief that with time and skill, it is possible and useful to determine an approximate range of developmental levels for almost any child (Schopler & Reichler, 1971). However, often no test is perfectly appropriate for the assessment of a young child with autism. Few tests employ methods appropriate for children with significant social and communication deficits (Parks, 1983). Few tests provide well-standardized and validated norms for children of both the relevant chronological ages and cognitive levels.

In the present paper, follow-up data were used to evaluate the relationship between stability of scores over a 5-year period and three factors in autistic children: age at initial assessment, severity of delay, and type of test. A previous study (Lord & Schopler, 1989) compared a subset of the preschool-age autistic subjects with a matched group of behavior-disordered communication-handicapped children. Results indicated that chronological age at first assessment had a significant effect on predictability of IQ at follow-up. However, the sample was too small to evaluate this finding in detail. For the present paper, a relatively large sample of autistic children made it possible for us to examine the relationship of chronological age, severity of delay, and type of test to each other and to changes in scores over time from several perspectives. In conjunction with findings from prospective studies (Freeman et al., 1985) and well-controlled follow-up studies (DeMyer et al., 1973, 1974; Rutter & Lockyer, 1967; Lockyer & Rutter, 1969, 1970), our results could be used to identify clinical and research implications for the use of certain tests in the early assessment of children with autism.

This report is based on analyses of scores from a relatively large number of children aged 2 and 3 years (i.e., 72) and 4 and 5 years (i.e., 70) at initial assessment, as well as an equal-size group of children initially aged 6 or 7 years. Of particular interest were two quite different tests (Parks, 1983), the Bayley Scales of Mental Development (Bayley, 1969) and the Merrill-Palmer Scale of Mental Tests (Stutsman, 1931) that were used in nearly all of the assessments of children under age 5. Though the tests differ in many ways, it is possible to administer both of them to young children who are unable to cope with school-like procedures that are built into the standardization of most tests intended for older and/or more advanced children (Terman & Merrill, 1973; Wechsler, 1974).

In this study, autistic children's scores on these tests are related to scores at follow-up on the same tests, the Leiter International Performance Scale (Arthur, 1952) and the performance score of the Wechsler Intelligence Scale for Children–Revised version (WISC-R; Wechsler, 1974). The objective is to assess the relationship between predictability and stability of early DQ/IQ scores and age at initial assessment, severity of delay, and type of test. This goal is approached in three ways. First, means on DQ/IQ are compared across groups categorized by age, severity of delay, and test. Second, the extent to which these variables accounted for performance IQ at follow-up is assessed. Third, patterns of change across time and test for individuals are compared.

Because the study used scores from assessments performed as part of an ongoing clinical service, tests were not randomly assigned to subjects. Children were given the most chronologically age-appropriate test on which they could receive a basal score. This constraint limits interpretation of our results because chronological age, cognitive level (and therefore, severity of delay), and choice of test are not independent factors. However, their confounding in this clinical sample provides an opportunity to assess some of the selection biases inherent in the intellectual assessment of any group of autistic children who cover more than the most narrow age and/or ability range.

Within the constraints of having to choose from what had been administered, we randomly selected one early and one later assessment per child. Often we had results from more than one developmental or intelligence test for a child within a period of a year or two. By making additional comparisons between tests given to the same child within the same period, it was possible, to some extent, to separate the effects of child-related characteristics, such as age at assessment, from test characteristics. In addition, scores on the Vineland Social Maturity Scale (Doll, 1965) were available for all children. These results are discussed, along with data concerning receptive vocabulary, in a separate paper (Lord & Schopler, 1988).

METHOD

Subjects

Out of a sample of over 1,000 children with communication disorders assessed at the TEACCH (Treatment and Evaluation of Autistic and Communication-handicapped Children) clinic betwen the ages of 2 and 7 years, 216 autistic children (171 male, 45 female) were identified who had received a complete intellectual assessment at age 7 years or younger and had been reassessed a minimum of 24 months later. Only those children who received DQ/performance IQ scores between 30 and 105 at initial assessment

on specific tests listed below were included. A more complete description of specific inclusion and exclusion factors is available in Lord and Schopler (1989), a report of a comparison of scores from 71 of the 213 autistic children included in the analyses described here with scores from 71 nonautistic children.

Children were classified as autistic on the basis of concordant clinical impressions of two psychologists using a Childhood Autism Rating Scale (CARS; Schopler, Reichler, & Renner, 1986) score of over 30 and, at the initial assessment, Rutter (1978) criteria. All but three children (all male) were classified as autistic during the later assessment using the same criteria. In addition to the 216 children with initial diagnoses of autism, there were also 5 children who were not identified as autistic initially but who, at the later assessment, received clinical diagnoses of autism and appropriate CARS scores. However, in order to maintain the most homogeneous sample, only children diagnosed as autistic during both initial and later assessments were included in this sample.

Of the 213 remaining autistic children (168 male, 45 female), 126 were white, 79 black, and 8 of other races. Race was confounded with social class. However, there were no differences in social class or racial distribution across sex, age, or IQ grouping, nor for sex across age or IQ groupings within the autistic sample.

As shown in Table I, longitudinal comparisons of three separate groups of autistic children were made. No child was included in more than one of these comparisons. The first comparison was between scores for children with a mean age (years-months) of 3-2 years (2-0 to 3-11) at initial assessment and 7-8 years (6-0 to 8-11) at follow-up. The second comparison was of scores for children aged 4-7 years (4-0 to 5-11) at initial assessment and 9-1 years (8-0 to 10-11) at follow-up. The third comparison was between scores of children assessed at age 6-6 years (6-0 to 7-11) and 11-1 years (10-0 to 12-11). Seventy-two subjects were included in the 3- to 7-year longitudinal comparison, 70 subjects in the 4- to 9-year comparison, and 71 subjects in the 6- to 11-year comparison. For each of these analyses, the mean time difference between the assessments was between 54 and 56 months.

Analyses were originally carried out both for longitudinal comparisons in which the time difference was fixed across scores (as described above) and longitudinal comparisons in which initial age at evaluation and interval between assessments were varied, while age at last evaluation was held constant at 10-12 years. Generally, the results were similar. However, the interval between assessment (when it varied from 2 to 10 years) sometimes had significant effects. In contrast, when the interval between assessments was held constant, age at later evaluation did not show any effect in any of the analyses. Because fixing the age at the follow-up assessment decreased the sample size, longitudinal comparisons were made across samples that had comparable intervals *between* assessments rather than the same age at final assessment.

Table I. Relationship Between IQ/DQ Scores at Preschool Age and School Age

Mean age (years–months)			IQ/DQ			Absolute IQ difference		Median IQ difference
Time 1	Time 2	*n*	Time 1	Time 2	*r*	*M*	*SD*	
3–2	7–8	72	56.64	63.52	.68	15.33	16.77	12
4–7	9–1	70	58.18	60.90	.81	12.18	8.71	11
6–6	11–1	71	58.39	57.65	.83	11.10	8.72	9

Measures

A detailed description of the measures and decisions for inclusion and exclusion of tests is available in Lord and Schopler (1989). Children in the youngest age groups were assessed using either the Bayley Infant Scales of Mental Development (Bayley, 1969) or the Merrill-Palmer Scale of Mental Yests (Stutsman, 1931). For the later assessments, children received the Merrill-Palmer, the Leiter International Performance Scale (Arthur, 1952), or the performance scale of the WISC-R (Wechsler, 1974). For some analyses, individual subjects were categorized according to levels of adaptive functioning as measured on the Vineland Social Maturity Scale: severe retardation (SQ = 30–49), mild retardation (SQ = 50–69), or nonretarded (SQ $\geq$ 70). To minimize confusion, results grouped in this way are described as grouped by adaptive level. Categorical adaptive level was used as an independent variable in this case so that IQ could be used as a dependent variable with sufficiently equivalent variability across time to permit use of parametric statistics. For nonparametric analyses, subjects were grouped according to IQ range (IQ = 30–49, 50–69, 70–105).

If more than one complete assessment was available for a child, an initial age was selected using a random numbers table. When statistical comparisons were made of correlations between tests given at different times, subjects with scores on more than one test were included in any appropriate comparisons of means across time. However, when comparisons were made across tests, scores from only one pair of assessments were randomly selected so that each subject was represented in only one set of correlations.

RESULTS

Preliminary Analyses

Preliminary analyses (ANOVAs with time as a repeated measure) were carried out for the longitudinal comparisons for each of the three groups.

Results indicated no main effects or second-order interactions for sex or race. Consequently, sex and race were dropped from further analyses. It is worth noting that females were nonsignificantly lower than males on IQ measures over time for all different tests and all samples. This general effect was reported earlier for overlapping samples (Lord, Schopler, & Revecki, 1982; Lord & Schopler, 1985).

Description of Statistical Analyses

Three types of statistical analyses were performed. ANOVAs were used to assess the effect on group means of major variables such as test, adaptive level, and the interaction between these variables and time. Multiple regressions were used to indicate the extent to which different variables (including IQ at initial assessment) accounted for the variance of IQ at follow-up. Chi-square tests were used to assess differences in the patterns of scores of individual subjects over time and/or test. Because of the large number of different analyses, rather than reporting results according to the type of statistic, results are reported according to each of the major variables in turn.

Effect of Age on Stability and Predictability of IQ

A 2(Time) × 3(Adaptive Level) × 3(Age) repeated measures ANOVA was performed on the combined samples of 213 autistic subjects. This analysis yielded main effects for adaptive level, simply confirming that grouping by Adaptive Level (i.e., SQ) resulted in mean differences in IQ. A Time × Age interaction occurred, $F(2, 207) = 10.88$, $p < .01$. Most of this effect was due to changes in IQ occurring for the 3- to 7-year group. Further analyses of the effect are described below as part of the discussion of the three-way interaction between time, age, and adaptive level.

A one-way ANOVA yielded a significant effect of age on the mean absolute difference in IQs for individual subjects from initial assessment to follow-up, $F(1, 211) = 8.76$, $p < .01$. As shown in Table I, mean difference scores decreased as age increased, from 15.33 IQ points for the group aged 3 to 7 years, to 12.18 points for the group aged 4 to 9 years, to 11.10 points for the group aged 6 to 11 years. Specific comparisons (Scheffé, 1953) between all pairs of age groups were significant, $ts > 3.0$, $ps < .05$. Overall, performance IQ/DQ scores over the follow-up intervals were quite stable, with stability increasing over the preschool years. As indicated earlier, identical analyses that held constant age at last testing yielded similar results. Thus, the effect appeared to be due to differences in age at initial testing rather than age at last assessment.

A stepwise multiple regression was performed separately for longitudinal comparisons of each group in order to assess the extent to which initial assessment variables of IQ and age (within the 2- to 3-year spans) predicted IQ at follow-up. At all three ages, initial IQ was the best predictor of IQ at follow-up [3 to 7 years: $R^2 = .68$, $F(2, 70) = 10.04$; 4 to 9 years: $R^2 = .81$, $F(2, 68) = 32.67$; 6 to 11 years: $R^2 = .83$, $F(2, 69) = 25.73$, $ps < .01$]. Age (within the 2-year spans) did not add significantly to the variance accounted for by initial IQ. Differences in correlations between initial IQ and follow-up IQ for the three age groups were not significant.

Effects of Adaptive Level/IQ Range at Initial Assessment on Stability and Predictability of Later IQ

The Time × Age × Adaptive Level ANOVA discussed earlier also yielded two significant two-way interactions: Time × Adaptive Level, $F(2, 207) = 9.62$, $p < .01$, and Age × Adaptive Level, $F(2, 207) = 7.60$, $p < .01$. There was also a three-way interaction for Time × Adaptive Level × Age, $F(2, 207) = 6.22$, $p < .01$. These effects reflected differences in direction and magnitude of IQ change over time for children of different adaptive levels at different ages. Specific comparisons (Scheffé, 1953) yielded only two significant changes over time out of the nine cells generated by the Age × Adaptive Level interaction, $ts > 2.58$; $ps < .05$. Mean IQ increased for severely retarded children from 38.20 ($SD = 11.65$) at age 3 to 56.68 ($SD = 11.94$) at age 7, and decreased for nonretarded children from 81.20 ($SD = 10.81$) at age 6 to 74.05 ($SD = 10.81$) at age 11.

More clinically relevant questions concern the direction and magnitude of changes in performance IQ over time for individual subjects who scored within different IQ ranges or adaptive levels at initial assessment. Separate chi-square tests were performed for subjects in each longitudinal comparison in order to observe the relationship between performance IQ at initial assessment and direction of change (increase or decrease) of IQ over time. As shown in Table II, IQs of many of the youngest children (i.e., 2- and 3-year-olds) increased (often only slightly) by the time they were 7 years old. In contrast, IQs of a greater number of the older children (i.e., 6- and 7-year-olds) decreased by the time they were 11 years old.

The different patterns in direction of change for different age groups and severity levels were even more apparent when only changes of 20 or more points were considered. Of 72 children aged 3 to 7 years, all nine changes of 20 points or greater were increases and all occurred within the subset of 22 children initially categorized as severely retarded. For the 70 children aged 4 to 9 years, eight changes of 20 points occurred. These changes were distributed across all three initial IQ groups and occurred in both directions

Table II. Number of Children in Each Longitudinal Group Showing Increases or Decreases in IQ Over Time[a]

	Follow-up IQ higher	Follow-up IQ lower	$\chi^2(2)$	n
3- to 7-year comparison				
Severe	21	1	15.04[c]	72
Mild	26	6		
Nonretarded	8	10		
4- to 9-year comparison				
Severe	19	6	14.40[c]	70
Mild	12	16		
Nonretarded	3	14		
6- to 11-year comparison				
Severe	6	9	6.41[b]	71
Mild	20	23		
Nonretarded	1	12		

[a]Subjects are categorized according to performance IQ range at initial assessment: Severe = 30–49, mild = 50-69, nonretarded = 70-105.
[b]$p < .05$.
[c]$p < .001$.

with about the same frequency. For the 71 children aged 6 to 11 years, all nine changes of 20 or more points were decreases. They were also distributed relatively equally across IQ groups.

Table III shows the distributions of the entire sample of 213 subjects according to IQ range at both initial and follow-up assessments. This distribution was not random, χ^2 (4, N = 213) = 38.82, $p < .01$. The greatest proportion of children remained within the same range for both assessments. However, a substantial minority of children changed IQ ranges.

Effect of Tests on Absolute IQ Scores, Stability, and Predictability

Initial Test

Repeated measure ANOVAs assessing the effect of Initial Test × Time and Later Test × Time were run separately for longitudinal comparisons of each of the three age groups. Main effects of Initial Test occurred at the two younger age levels [for age 3 to 7 years: $F(1, 70) = 19.54$; for age 4 to 9 years: $F(1, 68) = 4.72$, $ps < .05$]. Mean Bayley scores were consistently 10 or more points below Merrill-Palmer scores (and Leiter scores in the case of follow-up for the older age group) for both comparisons.

Time × Test interactions occurred for longitudinal comparisons for the two younger age groups as well [age 3 to 7 years: $F(1, 70) = 6.75$, $p < .01$; age 4 to 9 years: $F(1, 68) = 4.51$, $p < .05$]. For these two younger groups,

Table III. Distribution of Combined Samples of Autistic Children According to IQ Scores at First Assessment and at Follow-Up[a]

IQ range at first assessment	IQ range at follow-up			
	Severe	Mild	Nonretarded	Total
Severe	31	24	7	62
Mild	27	40	36	103
Nonretarded	2	10	36	48
Totals	60	74	79	213

[a]Subjects are categorized according to performance IQ range at initial and later assessments: Severe = 30–49, mild = 50–69, nonretarded = 70–105. These scores reflect the constraints on IQ at initial assessment used to select subjects. Children were still included in the later assessments if scores fell beyond these ranges (i.e., <30 or >105).

interactions reflected significant increases over time in performance IQ for children who had been given Bayleys, in contrast to no significant changes for children who had initially received Merrill-Palmers or Leiters, $ts > 5.00$, $ps < .05$. For example, the mean score for children who received a Bayley at age 3 was 39.91, compared to the same children's mean score of 53.67 achieved at age 7 on the later tests (i.e., Merrill-Palmer, Leiter, or WISC-R performance scale). In contrast, the mean score for 3-year-olds given the Merrill-Palmer at initial assessment was 64.30, compared to their mean score of 66.75 at age 7 on the three performance tests.

Chi-square tests were also employed within the different age groups to look at differences in the proportion of subjects whose IQs increased or decreased over time according to different tests. For all age groups, children who received Bayleys initially were more likely to show increases in performance IQ with increasing age, whereas children who received Merrill-Palmers were more likely to show decreases as they grew older, $\chi^2(2)$'s = 91.98, 13.42, 14.74, respectively, with increasing age, $ps < .005$.

It is important to remember that these groups did not comprise the same children. Children who were given the Bayley were, on the average, 6 months younger than children given the Merrill-Palmer. In addition, on performance tests at follow-up, the scores of children first given the Bayley were consistently lower for all the three samples than the scores of children initially assessed with the Merrill-Palmer.

However, it was possible to identify, out of the combined sample, 34 children who received both the Bayley and the Merrill-Palmer within a year (again, in all cases the Merrill-Palmer was given later, with a mean time difference of 6 months). As shown in Table IV, all or almost all children under age 5 at initial assessment, including both those assessed on the different tests

Table IV. Number of Two- to Five-Year-Old Children Showing Increases or Decreases When Tested Twice on the Same or Different Tests

	Secont test score higher	Second test score lower	Mean absolute IQ difference
Different tests			
Bayley/Merrill Palmer			
Given within 1 year	34	0	9.22
Given more than 2 years apart	53	4	13.64
Same test within 2 years			
Bayley/Bayley	15	12	7.63
Merrill-Palmer/Merrill-Palmer	30	34	4.21

within a year and those whose assessments were separated by more than 2 years, received higher scores on the Merrill-Palmer than the Bayley. In contrast, the numbers of increases and decreases were about equal for children who received either the Bayley or the Merrill-Palmer twice within 2 years (i.e., mean time diferences were between 20 and 22 months). Thus, differences between administration of the same test twice within 2 years were always smaller than differences between the Merrill-Palmer and Bayley, even when the period between assessments was shorter for the different-test than the same-test comparisons. Furthermore, the consistently lower follow-up scores of the Bayley recipients indicated a selection effect; lower-functioning children constituted a greater proportion of the Bayley samples than the early Merrill-Palmer samples.

Later Test

The only longitudinal group for which significant effects were found for Later Test was children in the age 6- to 11-year comparison. A main effect of Later Test, $F(2, 65) = 7.97$, $p < .01$, and a Later Test $\times$ Time interaction were found, $F(2, 65) = 4.38$, $p < .05$. Only specific comparisons (Scheffé, 1953) significant at $p < .05$ or less are reported below.

On the average, children given WISC-Rs at follow-up had higher performance IQs during both assessments (Time 1 $M = 74.37$; Time 2 $M = 69.64$) than children who were given Merrill-Palmers for the second assessment (Time 1 $M = 40.06$; Time 2 $M = 38.97$). These results were likely to have been due to selection, rather than to the particular tests, since the scores differed on the early assessments as well as on the later ones.

Children given Leiters at the second assessment had lower performance IQs for the initial assessments (Time 1 $M = 59.40$; Time 2 $M = 63.97$) than children who were given WISC-Rs at follow-up and higher performance IQs than children who continued to be given Merrill-Palmers. For their second

set of scores, children given the Leiter had higher scores on the average than children given the Merrill-Palmer, but were not significantly different from children given the WISC-R performance scale. Patterns for samples for the age groups 3 to 7 years and 4 to 9 years were similar to those of the older group, but nonsignificant.

Chi-square tests performed separately for each longitudinal sample were used to compare the number of children administered the Merrill-Palmer, Leiter, or WISC-R at follow-up who showed increases or decreases in performance IQ over time. Similar patterns of results were significant for the two younger samples, but not for the oldest group of children. For the age 3 to 7 group, children who were given Merrill-Palmers and Leiters as their second test tended to show higher performance IQs on their second assessment than their first, whereas children who were given WISC-Rs were equally split between increases and decreases, χ^2 (2, N = 72) = 5.99, $p < .05$. For comparisons of the age 4- to 9-year group, more children who were given the Leiter as their second test showed increases, and more children given the Merrill-Palmer showed decreases than children who were given the WISC-R, who were equally split upward and downward, χ^2 (2, N = 70) = 14.74, $p < .05$. However, it is important to remember that less able children generally were given Leiters and Merrill-Palmers rather than WISC-Rs. Mean decreases would have been more frequent had all children been given a WISC-R.

Table V presents correlations combined across all three age groups, for all children who received any pair of tests with assessments separated by more than 2 years. All correlations between different tests were statistically significant. Correlations between repeated administrations of the same tests were significantly, zs > 1.96, greater in all cases than correlations between two different tests (for the comparison of correlations, subjects were randomly assigned to one test-pair only; otherwise, all scores available were used). Changes of greater than 20 points were extremely rare when the same test was given twice. Mean absolute differences on repeated administration of the same test given 2 to 6 years later averaged less than 7 points for each of the comparisons in Table V, in contrast to mean differences of 11 to 15 points when two different tests were used.

In order to clarify if results described earlier for direction and magnitude were due to changes in tests or changes over time, scores were also compared for children who received two different tests within a 1-year period between the ages of 6 and 12 years (M interval = 7–9 months).

For 33 children given the Merrill-Palmer or Leiter and a WISC-R within the same 1-year period, which test socre was higher was equally divided when WISC-R performance scales were used. Absolute differences averaged less than 9 points. As in Shah and Holmes (1985), when full-scale WISC-R scores were used, these scores were almost always lower than Leiter or Merrill-Palmer scores.

Table V. Relationship Between IQ Scores for Assessments Separated by Two Years or More on the Same or Different Tests[a]

		M age (years–months)		
	n	Time 1	Time 2	*r*
Different tests				
Bayley/Merrill-Palmer	126	3–2	6–4	.53
Merrill-Palmer/Leiter	168	5–0	8–7	.59
Merrill-Palmer/WISC-R	31	5–3	9–3	.55
Leiter/WISC-R	25	9–0	11–11	.51
Same tests				
Bayley/Bayley	42	3–8	5–9	.75
Merrill-Palmer/Merrill-Palmer	137	4–11	8–2	.78
Leiter/Leiter	60	6–5	11–10	.86

[a]These samples include additional assessments of the 213 subjects to those reported for earlier, separate longitudinal comparisons. All correlations are significant at $p < .01$. Correlations between scores on the same test are all significantly higher than correlations between scores from different tests.

What was potentially more important then differences in IQ points was the marked selection effect determining which children were given which tests. Out of 213 children, 205 at some point received a Merrill-Palmer and 188 received a Leiter, but only 43 ever received a WISC-R performance scale. This last group of children had higher scores on all early tests, compared to children never given a WISC-R.

DISCUSSION

Good stability and predictability were found for intelligence and developmental quotients for groups of autistic children from preschool to school years, even including children assessed at 3 years and under. However, closer inspection reveals a number of exceptions to these findings.

First, like normally developing children (Hindley & Owen, 1978; Kopp & McCall, 1982; Sattler, 1982), young autistic children followed over 5 years show substantial changes in absolute scores as individuals. Unlike the children in the broader groupings of Freeman, Ritvo, and colleagues (Freeman et al., 1985), children in our longitudinal samples showed a significant amount of movement between IQ ranges. For example, over one quarter of the children who scored between 51 and 69 on an early assessment scored below 50 in a later assessment. However, most changes were between contiguous ranges and within 10 to 12 points. Only two children who scored over 70 at 6 years or older had scored below 50 at their earlier assessment.

These findings suggest that IQ is a relatively stable characteristic even of quite young autistic children. However, narrow categorization of individual children by early DQ/performance IQ scores is not appropriate. Fine distinctions based on IQ differences of 10–15 points (such as used by some school systems to differentiate "slow learners" from "learning-disabled children of normal intelligence," or between "mild" and "moderate" retardation) would not have been reliable for the autistic children studied here.

Second, large increases occurred almost exclusively in one particular group, children 3 years or under whose early scores on the Bayley Scale of Mental Development were below 50 points. It is important to note that, even when these children showed significant gains, 20 out of 22 of them continued to score in the range of mental retardation. However, as assessed between ages 6 and 12, just under half of these children were no longer classified as severely mentally retarded, even though they remained mentally handicapped and autistic. These results cannot be accounted for by regression to the mean, which would have yielded changes of 6 to 8 points across all age groups for the lowest ranges (Furby, 1973).

One likely source of this variability is the difference between the Bayley and the performance tests, particularly the Merrill-Palmer. The tests share the advantages of having many visually interesting materials, allowing modification of the order of items, and including in the protocol the demonstration of many tasks. However, the Bayley and Merrill-Palmer are also quite different from each other in content and construction. On the Bayley, many "social" behaviors, such as smiling and attending to the examiner's actions, are scored. Language items become more frequent as age increases. In contrast, the Merrill-Palmer primarily involves the manipulation of objects without social referents. Language items represent proportionately fewer scores with increasing "mental age"; these items can be completely excluded by treating them as refusals. Thus, variability over time in absolute scores and in predictability for other tests may lie in differences between the tests themselves as well as in characteristics of the children being tested.

Results for high-functioning autistic children followed a different pattern than results for children with lower IQs. Even though, across age, children who scored above 70 on early assessments were the most likely to show decreases in IQ, they were also the least likely of any of the groups of children to shift IQ range. The latter finding was due in part to the size of the range but also to lower absolute variability within this group. This result supports earlier findings (DeMyer et al., 1973; Freeman et al., 1985; Lockyer & Rutter, 1969). The decreases in IQ seem most likely to be a function of the use of the WISC-R as the later test, compared with initial scores from either the Merrill-Palmer or the Leiter. In this study, 75% of the children who scored as nonretarded in preschool scored above 70 in performance IQ at age 7 or older.

On the whole, even for early assessments with the Bayley, correlations among tests were relatively high and similar to those found in previous studies (Lockyer & Rutter, 1970; Shah & Holmes, 1985). However, because of the clinical nature of the study, selection effects were clearly present. Findings indicated that the source of different patterns of results lay both in the tests themselves and in the selection of tests for each child. Age was one variable; suspected intellectual level was probably another. Only one of the 3-year-olds scoring below 50 was given the Merrill-Palmer; all others were given the Bayley. All of the 3-year-olds who scored above 70 did so on the Merrill-Palmer; children who scored between 51 and 69 were divided equally between the Bayley and the Merrill-Palmer.

The highest functioning children were given the most chronologically age-appropriate tests. Thus, children who continuted to be given the same preschool test when they were no longer preschool children, showed decreases in performance IQ and showed lower IQs at both testings than children able to move on to a test for an older age group. As for other populations (Hindley & Owen, 1978), predictability over time was highest when a child was given the same test twice. In our study, this situation was associated with having low scores on both the first and last test. The stability of scores above 70 was therefore particularly encouraging, because these children were most likely to have been assessed on different tests on the different occasions.

However, though the comparability for scores from the performance tests given at school age (i.e., Merrill-Palmer, Leiter, WISC-R) was relatively good for the higher functioning children (*M* IQs = 64–69) who were given two or three of the tests, we do not know what would have happened had the WISC-R been attempted with all the initial Merrill-Palmer and Leiter recipients. Some of the stability we found was likely due to the selective exclusion of children who would not have scored at all on the WISC-R. Thus, on the basis of our results, we cannot say that a Merrill-Palmer or a Leiter IQ of 50 is necessarily equivalent to a similar score on the WISC-R performance scale. More systematic, prospective study of this question is needed. In the meanwhile, in research for which IQ is considered an important factor, one must ensure that the bases (test and age) for its determination do not vary across comparison groups. That is, subjects matched on IQ should, in most cases, also be matched on type of test and chronological age at testing.

The issue of treatment effects needs to be addressed, though it was not a focus of this study. Only children who participated in at least 6 weeks of extended diagnostic, parent-as-therapist treatment were included in the study. Thus, differences in the stability and predictability of scores among age groups and in intellectual range adaptive levels were not due to general effects of treatment. On the other hand, the unusual pattern of change in the group of young, low-scoring children could be due in part to both direct and in-

direct treatment effects (Schopler, 1987). Many behaviors associated with better test performance, such as attention, cooperation, and motivation (Lord & Schopler, 1988; Sattler, 1982), are frequently targeted in the initial phases of treatment at the TEACCH program. Contexts for treatment are often very similar to those of the testing (e.g., using familiar materials in a structured way to work with an adult in a relatively confined space). The youngest and lowest scoring children may have had particular difficulty in these areas and hence showed the greatest response to this aspect of treatment.

Treatment may also have an indirect effect on test scores by giving children sufficient test-taking skills so that a structured performance test such as the Merrill-Palmer or the Leiter can be administered, instead of the more socially based but more flexibly administered Bayley. This research did not include an untreated control group, so test–treatment interactions cannot be rejected, though it seems likely the increases were part of more general changes in development (Lord & Schopler, 1988; Rutter, 1985). On the other hand, as for the school age performance tests, in any sort of group comparison, particularly of the results of early treatment programs, care must be taken that children are matched across group on age and type of test as well as on actual test score, or one may have difficulty determining if improvements are due to changes in test and test taking or to intervention (Lovaas, 1987).

Overall, some clinical implications are also clear. The usefulness of performance IQ as a stable summary measure of mental handicap is supported once again. Early DQs/performance IQs of less than 50 are reliable indicators of the presence of mental handicap in autism. Early IQs can be used to predict later intellectual functioning, but the age of the child and the test employed must be taken into account when evaluating the severity of delay. Changes in test scores must always be evaluated (for research or clinical purposes) with respect to the age of the child, differences in the tests employed, and the interval between assessments.

Ironically, findings that intellectual skills of some autistic children deteriorate at adolescence suggest that the very earliest assessments might be better predictors of adult functioning than are the more optimal scores achieved during school years (Gillberg & Steffenburg, 1987; Lockyer & Rutter, 1969; Waterhouse & Fein, 1984). It may be appropriate to *begin* to prepare parents of very young children for the possibility that an autistic child who scores below 50 at age 3 years will remain significantly mentally handicapped, as well as autistic. However, it would be wrong to assume that the child will necessarily continue to fall in the *severe* range of mental retardation. Reassessments of the preschool child particularly when the child can cope with a test other than the Bayley, are necessity before determination of level of mental retardation can be made in conjunction with a diagnosis of autism.

REFERENCES

Arthur, G. (1952). *The Arthur adaptation of the Leiter International Performance Scale.* Chicago: The Psychological Service Center Press.

Bayley, N. (1969). *Manual for the Bayley Scales of Infant Development.* New York: Psychological Corp.

DeMyer, M. K., Barton, S., Alpern, G. D., Kimberlin, C., Allen, J., Yang, E., & Steele, R. (1974). The measured intelligence of autistic children. *Journal of Autism and Childhood Schizophrenia, 4,* 42-60.

DeMyer, M. K., Barton, S., DeMyer, W. E., Norton, J. A., Allen, J., & Steel, R. (1973). Prognosis in autism: A follow-up study. *Journal of Autism and Childhood Schizophrenia 3,* 199-245.

Doll, E. A. (1965). *Vineland Social Maturity Scale.* Circle Pines, MN: American Guidance Service.

Freeman, B. J., Ritvo, E. R., Needleman, R., & Yokota, A. (1985). The stability of cognitive and linguistic parameters in autism: A five-year prospective study. *Journal of the American Academy of Child Psychiatry, 24,* 459-464.

Gillberg, C., & Steffenburg, S. (1987). Outcome and prognostic factors in infantile autism conditions: A population-based study of 46 cases followed through puberty. *Journal of Autism and Developmental Disorders, 17,* 273-288.

Hays, W. L. (1963). *Statistics for the social sciences.* New York: Holt, Rinehart & Winston.

Hindley, C. B., & Owen, C. F. (1978). The extent of individual changes in IQ for ages between 6 months and 17 years, in a British longitudinal sample. *Journal of Child Psychology and Psychiatry, 19,* 329-350.

Kopp, C. B., & McCall, R. B. (1982). Predicting later mental performance for normal, at risk, and handicapped infants. In P. B. Baltes & O. G. Brim, Jr. (Eds.), *Life-span development and behavior* (Vol. 4, pp. 33-61). New York: Academic Press.

Lockyer, L., & Rutter, M. (1970). A five to fifteen-year follow-up study of infantile psychosis: IV. Patterns of cognitive ability. *British Journal of Social and Clinical Psychology, 9,* 152-163.

Lord, C., & Schopler, E. (1985). Differences in sex ratios in autism as a function of measured intelligence. *Journal of Autism and Developmental Disorders, 15,* 185-193.

Lord, C., & Schopler, E. (1988). Intellectual and developmental assessment of autistic children from preschool to schoolage: Clinical implications of two follow-up studies. In E. Schopler & G. B. Mesibov (Eds.), *Diagnosis and assessment in autism* (pp. 167-181). New York: Plenum Press.

Lord, C., & Schopler, E. (1989). Stability of assessment results of autistic and nonautistic language-impaired children from preschool years to early school age. *Journal of Child Psychology and Psychiatry. 30,* 575-590.

Lord, C., Schopler, E., & Revicki, D. (1982). Sex differences in autism. *Journal of Autism and Developmental Disorders, 12,* 317-330.

Lovaas, O. I. (1987). Behavorial treatment and normal educational and intellectual functioning in young autsitic children. *Journal of Consulting and Clinical Psychology, 55,* 3-9.

Parks, S. L. (1983). The assessment of autistic children: A selective review of available instruments. *Journal of Autism and Developmental Disorders, 13,* 225-268.

Rutter, M. (1978). Diagnosis and definition of childhood autism. *Journal of Autism and Developmental Disorders, 8,* 139-161.

Rutter, M. (1985a). *A clinician's guide to child psychiatry* (pp. 48-78). New York: Free Press.

Rutter, M. (1985b). The treatment of autistic children. *Journal of Child Psychology and Psychiatry, 26,* 193-214.

Rutter, M. & Lockyer, L. (1967). A five to fifteen year follow-up study of infantile psychosis: I. Description of sample. *British Journal of Psychiatry, 113,* 1169-1182.

Sattler, J. M. (1982). *Assessment of children's intelligence and special abilities.* Boston: Allyn & Bacon.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika, 40,* 87-104.

Schopler, E. (1987). Specific and nonspecific factors in the effectiveness of a treatment system. *American Psychologist, 42,* 376-383.

Schopler, E., & Reichler, R. J. (1971). Problems in the developmental assessment of psychotic children. *Excerpta Medica International Congress Series, 274,* 1307-1311.

Schopler, E., Reichler, R. J., & Renner, B. R. (1986). *The Childhood Autism Rating Scale (CARS) for diagnostic screening and classification of autism.* New York: Irvington Publishers.

Shah, A., & Holmes, N. (1985). The use of the Leiter International Performance Scale with autistic children. *Journal of Autism and Developmental Disorders, 15,* 195-204.

Stutsman, R. (1931). Guide for administering the Merrill-Palmer Scale of Mental Tests. In L. M. Terman (Ed.), *Mental measurement of preschool children* (pp. 139-262). New York: Harcourt, Brace & World.

Terman, L. M., & Merrill, M. A. (1973). *Stanford-Binet Intelligence Scale Form L-M.* Boston: Houghton-Mifflin.

Waterhouse, L., & Fein, D. (1984). Developmental trends in cognitive skills for children diagnosed as autistic and schizophrenic. *Child Development, 55,* 236-248.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children-Revised.* New York: Psychological Corp.