

Long-Term Outcome for Children With Autism Who Received Early Intensive Behavioral Treatment

John J. McEachin, Tristram Smith, and O. Ivar Lovaas
University of California, Los Angeles

After a very intensive behavioral intervention, an experimental group of 19 preschool-age children with autism achieved less restrictive school placements and higher IQs than did a control group of 19 similar children by age 7 (Lovaas, 1987). The present study followed-up this finding by assessing subjects at a mean age of 11.5 years. Results showed that the experimental group preserved its gains over the control group. The 9 experimental subjects who had achieved the best outcomes at age 7 received particularly extensive evaluations indicating that 8 of them were indistinguishable from average children on tests of intelligence and adaptive behavior. Thus, behavioral treatment may produce long-lasting and significant gains for many young children with autism.

Infantile autism is a condition marked by severe impairment in intellectual, social, and emotional functioning. Its onset occurs in infancy, and the prognosis appears

This study was supported by Grant No. MH-11440 from the National Institute of Mental Health. The study was based on a dissertation submitted to the University of California, Los Angeles, Department of Psychology, in partial fulfillment of the requirements for the doctoral degree. The authors express their deep appreciation to the many students at UCLA who served as therapists and helped to make this study possible. Special thanks to Bruce Baker and Duane Buhrmester, who helped in the design of this study. Requests for reprints of this article, copies of the Clinical Rating Scale, or additional information about this study should be sent to O. Ivar Lovaas, 405 Hilgard Ave., UCLA, Department of Psychology, Los Angeles, CA 90024-1563.

to be extremely poor (Lotter, 1978). For example, in the longest prospective follow-up study with a sound methodological design, Rutter (1970) found that only 1 of 64 subjects with autism (fewer than 2%) could be considered free of clinically significant problems by adulthood, as evidenced by holding a job, living independently, and maintaining an active and age-appropriate social life. The remaining subjects showed numerous dysfunctions, such as marked oddities in behavior, social isolation, and florid psychopathology. The majority of subjects required supervised living conditions.

Professionals have attempted a wide variety of interventions in an effort to help children with autism. For many years, no scientific evidence showed that any of these interventions brightened the children's long-term prognosis (DeMyer et al., 1981). How-

ever, since the 1960s, one of these interventions, behavioral treatment, has appeared promising. Behavioral treatment has been found to increase adaptive behaviors such as language and social skills, while decreasing disruptive behaviors such as aggression (DeMyer, Hingtgen, & Jackson, 1981; Newsom & Rincover, 1989; Rutter, 1985). Furthermore, behavioral treatment has been continuously refined and improved as a result of ongoing research efforts at a number of sites (Lovaas & Smith, 1988).

Some recent evidence has indicated that behavioral treatment has developed to the point that it can produce substantial improvements in the overall functioning of young children with autism (Simeonsson, Olley, & Rosenthal, 1987). Lovaas (1987) provided approximately 40 hours per week of one-on-one behavioral treatment for a period of 2 years or more to an experimental group of 19 children with autism who were under 4 years of age. This intervention also included parent training and mainstreaming into regular preschool environments. When re-evaluated at a mean age of 7 years, subjects in the experimental group had gained an average of 20 IQ points and had made major advances in educational achievement. Nine of the 19 subjects completed first grade in regular (nonspecial education) classes entirely on their own and had IQs that increased to the average range. By contrast, two control groups totalling 40 children, also diagnosed as autistic and comparable to the experimental group at intake, did not fare nearly as well. Only one of the control subjects (2.5%) attained normal levels of intellectual and educational functioning.

These data suggest that behavioral treatment is effective. However, the durability of treatment gains is uncertain. In one prior major study, Lovaas, Koegel, Simmons, and Long (1973) found that children with autism regressed following the termination of treatment. Other studies have shown that children with autism may display increased difficulties when they enter adolescence (Kanner, 1971; Waterhouse & Fein, 1984).

Also, as was stated in the first follow-up (Lovaas, 1987), "Certain residual deficits may remain in the normal-functioning group that cannot be detected by teachers and parents and can only be isolated on closer psychological assessment, particularly as these children grow older" (p. 8). This possibility points to the need for a more detailed assessment and for continued follow-ups of the group over time.

The present investigation contained two parts: In the first part we examined whether several years after the evaluation at age 7, the experimental group in Lovaas's (1987) study had maintained its treatment gains. Subjects in the experimental group and one of the control groups completed standardized tests of intellectual and adaptive functioning. The groups were then contrasted with each other, and their current performance was compared to their performance on previous assessments. The second part of the investigation focused on those subjects who had achieved the best outcome at the end of first grade in the Lovaas (1987) study (i.e., the 9 subjects who were classified as normal functioning out of the 19 in the experimental group). We examined the extent to which these best-outcome subjects could be considered free of autistic symptomatology. A test battery was constructed to assess a variety of possible deficits: for example, idiosyncratic thought patterns, mannerisms, and interests; lack of close relationships with family and friends; difficulty in getting along with people; relative weaknesses in certain areas of cognitive functioning, such as abstract reasoning; not working up to ability in school; flatness of affect; absence or peculiarity in sense of humor. Possible strengths to be identified included normal intellectual functioning, good relationships with family members, ability to function independently, appropriate use of leisure time, and adequate socialization with peers. Numerous methodological precautions were taken to ensure objectivity of the follow-up examination.

Method

Subjects and Background

Characteristics of the subjects and their treatment have been described elsewhere (Lovaas, 1987) and will only be summarized here. The initial treatment study contained 38 children who, at the time of intake, were very young (less than 40 months if mute, less than 46 months if echolalic) and had received a diagnosis of autism from a licensed clinical psychologist or psychiatrist not involved in the study. These 38 subjects were divided into an experimental group and a control group. The assignment to groups was made on the basis of staff availability. At the beginning of each academic quarter, treatment teams were formed. The clinic director and staff members then determined whether any opening existed for intensive treatment. If so, the next referral received would enter the experimental group; otherwise, the subject entered the control group. The experimental group contained 19 children who received 40 or more hours per week of one-to-one behavioral treatment for 2 or more years. The control group was comprised of 19 children who received a much less intensive intervention (10 hours a week or less of one-to-one behavioral treatment in addition to a variety of treatments provided by community agencies, such as parent training or special education classes). The initial study also included a second control group, consisting of 21 children with autism who were followed over time by a nearby agency but who were never referred for this study. However, these 21 subjects were not available for the present investigation. On standardized measures of intelligence, the second control group did not differ from either the experimental group or the first control group at intake, nor did it differ from the first control group when evaluated again when the subjects were 7 years old. These findings suggest that, as measured by standardized tests, (a) the children with autism who were referred to us for

treatment were comparable to children with autism seen elsewhere and (b) the minimal treatment provided to the first control group did not alter intellectual functioning.

Statistical analysis of an extensive range of pretreatment measures confirmed that the experimental group and control group were comparable at intake and closely matched on such important variables as IQ and severity of disturbance. The mean chronological age (CA) at diagnosis for subjects in the experimental group was 32 months. Their mean IQ was 53 (range 30 to 82; all IQs are given as deviation scores). The mean CA of subjects in the control group was 35 months; their mean IQ was 46 (range 30 to 80). Most of the subjects were mute, all had gross deficiencies in receptive language, none played with peers or showed age-appropriate toy play, all were emotionally withdrawn, most had severe tantrums, and all showed extensive ritualistic and stereotyped (self-stimulatory) behaviors. Thus, they appeared to be a representative sample of children with autism (Lovaas, Smith, & McEachin, 1989). A more complete presentation of the intake data was reported by Lovaas (1987).

The children in the experimental group and control group received their respective treatments from trained student therapists who worked in the child's home. The parents also worked with their child, and they received extensive instruction and supervision on appropriate treatment techniques. Whenever possible, the children were integrated into regular preschools. The treatment focused primarily on developing language, increasing social behavior, and promoting cooperative play with peers along with independent and appropriate toy play. Concurrently, substantial efforts were directed at decreasing excessive rituals, tantrums, and aggressive behavior. (For a more detailed description of the intervention program, see the treatment manual [Lovaas et al., 1980] and instructional videotapes that supplement the manual [Lovaas & Leaf, 1981].)

At the time of the present follow-up (1984-1985), the mean CA of the experimen-

tal group children was 13 years (range = 9 to 19 years). All children who had achieved normal functioning by the age of 7 years had ended treatment by that point. (*Normal functioning* was operationally defined as scoring within the normal range on standardized intelligence tests and successfully completing first grade in a regular, nonspecial education class entirely on one's own.) On the other hand, some of the children who had not achieved normal functioning at 7 years of age had, at the request of their parents, remained in treatment. The length of time that experimental subjects had been out of treatment ranged from 0 to 12 years (mean = 5), with the normal-functioning children having been out for 3 to 9 years (mean = 5).

The mean age of subjects in the control group was 10 years (range 6 to 14). The length of time that these children had been out of treatment ranged from 0 to 9 years (mean = 3). Thus, experimental subjects tended to be older and had been out of treatment longer than had control subjects. This difference in age occurred because the first referrals for the study were all assigned to the experimental group due to the fact that referrals came slowly (7 in the first 3.5 years) and therapists were available to treat all of them. (As noted earlier, subjects were assigned to the experimental group if therapists were available to treat them; otherwise, they entered the control group.)

Statistical analyses were conducted to test whether a bias resulted from the tendency for the first referrals to go into the experimental group. For example, it is conceivable that the first referrals could have been higher functioning at intake or could have had a better prognosis than subsequent referrals. If so, the subject assignment procedure could have favored the experimental group. To assess this possibility, we correlated the order of referral with intake IQ and with IQ at the first follow-up (age 7 years). Pearson correlations were computed across both groups and within each group. These analyses indicated that the order in which subjects were referred was not associated

with intake IQ or outcome IQ. Consequently, although the tendency for the first referrals to enter the experimental group created a potential bias, the data indicate that this was unlikely.

Procedure

The assessment procedure included ascertaining school placement and administering three standardized tests. Information on school placement was obtained from subjects' parents, who classified their child as being in either a regular or a special education class (e.g., a class for children with autism or mental retardation, language delays, multihandicaps, or learning disabilities). The three standardized tests were as follows:

1. *Intelligence test.* The Wechsler Intelligence Scale for Children-Revised (Wechsler, 1974) was administered when subjects were able to provide verbal responses. This included all 9 best-outcome experimental subjects plus 8 of the remaining 10 experimental subjects and 6 of the 19 control subjects. Subjects who were not able to provide verbal responses, the Leiter International Performance Scale (Leiter, 1959) and the Peabody Picture Vocabulary Test-Revised (Dunn, 1959) were administered. All of these tests have been widely used for the assessment of intellectual functioning in children with autism (Short & Marcus, 1986).

2. *The Vineland Adaptive Behavior Scales* (Sparrow, Balla, & Cicchetti, 1991). The Vineland is a structured interview administered to parents assessing the extent to which their child exhibits behaviors that are needed to cope effectively with the everyday environment.

3. *The Personality Inventory for Children* (Wirt, Lachar, Klinck, & Seat, 1977). This measure is a 600-item true-false questionnaire filled out by parents that assesses the extent to which their children show various forms of psychological disturbance (e.g., anxiety, depression, hyperactivity, or psychotic behavior).

These three tests were intended to provide a comprehensive evaluation of intellectual, social, and emotional functioning. All of the tests have been standardized on average populations. Hence, they provide an objective basis for comparing subjects to children without handicaps across the various areas that they assess.

Data were obtained on all subjects except one girl in the control group, who was known to be institutionalized and functioning very poorly. The 9 best-outcome subjects (those who had been classified as normal functioning at age 7) received particularly extensive evaluations, as outlined later. Of the 28 remaining subjects, 17 were evaluated by staff members in our treatment program, and 11 received evaluations from outside agencies such as schools or psychology clinics. (In some cases, the outside agencies did not administer all of the measures in this battery.)

Evaluation of Best-Outcome Subjects.

To ensure objectivity in the evaluation of the best-outcome subjects, we arranged for blind administration and scoring of all tests for these subjects as follows. A psychologist not associated with the study recruited advanced graduate students in clinical psychology to administer the tests. The examiners were not familiar with the history of the children, and the psychologist told them simply that the testing was part of a research study on assessment of children. The psychologist advised them that the nature of the study necessitated providing only certain standard background information: age, school placement and grade, and parent's name and phone number. To increase the heterogeneity of the sample and to control for any examiner bias, each examiner also tested one or more subjects who were matched in age to the experimental subjects and had no history of behavioral disturbance. The examiners were randomly assigned an approximately equal number of subjects for testing in the experimental group and the comparison group. Two experimental subjects were not living in the local area. Therefore, for

each of them, the psychologist recruited a tester from the subject's hometown area as well as an age-matched control subject, and data were collected as just described. In addition, the child's examiner filled out a clinical rating scale following a structured interview that covered a list of standard topics, including friendships, family relations, and school and community activities. The interview was designed both for eliciting content and for sampling interpersonal style. The rating scale consisted of 22 items, each scored 0 (best clinical status) to 3 (marked deviance) points. The items were designed to include likely areas of difficulty for children with autism of average intelligence (e.g., compulsive or ritualistic behavior, empathy for and interest in others, a sense of humor) as well as areas of potential difficulty for the general child population (e.g., depressed mood, anxiety, hyperactivity). (The complete scale and a copy of instructions for the clinical interview can be obtained by writing to the third author).

Results

Experimental Versus Control Group

This first section examines the overall effects of treatment through comparison of the follow-up data from the 19 subjects who received the intensive (experimental) treatment to the data from those who received the minimal (control) treatment. Data were obtained from all subjects on school placement and from all but one subject in the control group on IQ. On the Vineland, scores were obtained for 18 of 19 experimental subjects and 15 of 19 control subjects. The lowest availability of follow-up scores was on the Personality Inventory for Children, with scores for 15 experimental subjects and 12 control subjects.

The subjects in the control group who had Personality Inventory for Children scores did not appear to differ from subjects who were missing these scores, as compared on

t tests for differences in intake IQ, IQ at 7 years old, or IQ in the present study.

As noted earlier, 17 of the 29 subjects who were not in the best-outcome group were evaluated by Project staff members, 11 were evaluated by outside agencies, and 1 was not evaluated. To check whether Project staff members were biased in their evaluations or in their selection of which subjects to evaluate, we used *t* tests to compare subjects they evaluated to those evaluated by outside agencies on intake IQ, IQ at age 7 years, and IQ in the present study. No significant differences between subjects evaluated by Project staff members and those evaluated by outside agencies were found.

School Placement. In the experimental group, 1 of the 9 subjects from the best-outcome group who had attended a regular class at age 7 (J. L.) was now in a special education class. However, 1 of the other 10 subjects had gone from a special education class to a regular class and was enrolled in a junior college at the time of this follow-up. The remaining experimental subjects had not changed their classification. Overall, then, the proportion of experimental subjects in regular classes did not change from the age 7 evaluation (9 of 19, or 47%). In the control group, none of the 19 children were in a regular class, as had been true at the age 7 evaluation. The difference in classroom placement between the experimental group and the control group was statistically significant, $\chi^2(1, N = 38) = 19.05, p < .05$.

Intellectual Functioning. The test scores for the experimental group and control group on intellectual functioning, adaptive and maladaptive behaviors, and personality functioning are summarized in Table 1. As can be seen in the table, the experimental group at follow-up had a significantly higher mean IQ than did the control group. This difference was significant, $t(35) = 2.97, p < .01$. Eleven subjects (58%) in the experimental group obtained Full-Scale IQs of at least 80; only 3 subjects (17%) in the control group did as well. The scores were similar to those obtained by the experimental group and con-

trol group at age 7 (mean IQs of 83 and 5 respectively), indicating that the experimental group had maintained its gains in intellectual functioning between age 7 and the time of the current evaluation.

Table 1
Mean Scores and SDs by Group and Measure at Follow-Up

Measure	Group		Control
	Experimental	Control	
	Mean	SD	Mean
IQ	84.5	32.4	29.3
Vineland*			28.3
Communication	5.1	28.4	25.4
Daily Living Skills	73.1	26.9	18.9
Socialization	75.5	26.8	
Adaptive Behavior			
Composite	71.6	26.8	21.3
Maladaptive Behavior	10.6	8.2	7.2
PIC ^b Scales			
Mean elevation	61.8	10.2	8.1
Scales > 70	4.0	3.9	2.8

*Vineland Adaptive Behavior Scale. ^bPersonality Inventory for Children.

Adaptive and Maladaptive Behavior. On the Vineland, the mean overall or Composite score was 72 in the experimental group and 48 in the control group. (The average score for the general population on this test is 100, with a standard deviation [SD] of 15.) On the three subscales—Communication, Daily Living, and Socialization—each score closely paralleled the Composite score. The interaction between the groups and the subscales was not significant, indicating that across the three subscales, the experimental group consistently scored higher than did the control group. As can be seen in Table 1, Maladaptive Behavior was significantly higher in the control group, $t(31) = 2.39, p < .05$. The mean score for the control group was in the clinically significant range whereas that of the experimental group was not. (Scores of 13 and above are considered to be indicative of clinically significant levels of maladaptive behavior at ages 6 to 9 years; 12 or above, at 12 to 13 years; and 10 or above, at 14 years and older.) Thus, the findings indicate that the experimental group showed more adaptive behaviors and fewer maladaptive behav-

iors than did the control group.

Personality Functioning. Scores for the experimental group and control group did not differ on overall scale elevation, with mean t scores of 62 and 65, respectively. (On this test, the mean t score for the general population is approximately 50 [$SD = 10$].) T scores above 60 are considered indicative of possible or mild deviance, whereas t scores above 70 are viewed as suggesting a clinically significant problem, namely, one that may require professional attention. There was a significant interaction between the groups and the individual scales on this test, $F(15, 390) = 2.36, p < .01$. Results of the Tukey test indicated that the most reliable difference between groups occurred on the Psychosis scale, on which the experimental subjects had a mean of 78 and the control subjects had a mean of 104, $F(1, 26) = 8.53, p < .01$. Seven subjects in the experimental group scored in the clinically preferred range (below 70), whereas no subjects in the control group scored that low. Only one other scale showed a significant difference, Somatic Concerns, $F(1, 26) = 4.60, p < .05$. The control subjects tended to display a below average level of somatic complaints (mean of 45 as compared to 54 for the experimental subjects).

Best-Outcome Versus Nonclinical Comparison Group

A t test indicated no significant difference in age between the best-outcome group and the comparison group of children without a history of clinically significant behavioral disturbance. Subjects in the best-outcome group had a mean age of 12.42 years (range 10.0 to 16.25) versus 12.92 years (range 9.0 to 15.17) for the nonclinical comparison group. Scores on the WISC-R and clinical rating scale were obtained for all subjects; 1 experimental subject and 2 nonclinical comparison subjects were missing Vineland scores, and 2 experimental subjects and 1 nonclinical comparison subject were missing Personality Inventory for

Children scores. Both the Vineland and Personality Inventory for Children were completed by parents. In cases where these scores were not obtained, the parents had declined to participate.

On the measures that provide standardized scores, the functioning of the best-outcome subjects was measured most precisely by comparing the best-outcome group against the test norms. Therefore, this analysis is of primary interest. Data for the nonclinical comparison group are mainly useful in confirming that the assessment procedures were valid and in providing a contrast group for the one measure without norms, the Clinical Rating Scale. For the nonclinical comparison group, it will suffice to summarize the results as follows: On the WISC-R this group had mean IQs of 116 Verbal, 118 Performance, and 119 Full-Scale. On the Vineland the group obtained mean standard scores of 102 Communication, 100 Daily Living Skills, 102 Socialization, and 101 Composite. The mean scale score on the Personality Inventory for Children was 49. Thus, the nonclinical comparison group displayed above-average or average functioning across all areas that were assessed.

The next section is focused on the functioning of the best-outcome group on IQ, adaptive and maladaptive behavior, and personality measures and contrasts the best-outcome subjects with the comparison subjects on the Clinical Rating Scale.

Intellectual Functioning. Table 2 presents the IQ data for each subject in the best-outcome group and the mean scores for the group. This table shows that, as a whole, the 9 best-outcome subjects performed well on the WISC-R. Their IQs placed them in the high end of the normal range, about two thirds of an SD above the mean. Their Full-Scale IQs ranged from 99 to 136.

Subjects' scores were evenly distributed across a range from 80 to 125 on Verbal IQ and from 88 to 138 on Performance IQ. The subjects averaged 3 points higher on Performance IQ than Verbal IQ. Two of them (J. L. and A. G.) had at least a 20-point difference

Table 2
WISC-R Scores of the Best-Outcome Subjects

Subject	Verbal					Performance					WISC-R IQ		
	Infrm	Simil	Arith	Vocab	Compr	PicC	PicA	BlkD	ObjA	Cod	VIQ	PIQ	Full
R.S.	12	12	13	9	11	10	9	13	12	11	106	106	106
M.C.	17	19	11	14	10	12	16	19	19	11	125	138	136
M.M.	14	13	10	14	11	12	11	11	11	8	114	102	109
L.B.	12	16	11	13	15	7	12	17	17	19	119	131	128
J.L.	6	9	7	4	8	18	11	16	14	7	80	123	100
D.E.	9	17	8	10	15	13	9	12	9	17	98	114	105
A.G.	7	14	12	11	13	9	4	8	11	10	108	88	99
B.W.	12	11	10	10	9	7	10	9	11	10	102	95	99
B.R.	11	14	11	13	16	12	10	12	11	10	118	108	114
Mean	11.1	13.9	10.3	10.9	12.0	11.1	10.0	12.0	11.0	10.0	111.1	111.1	111.1

Note. Infrm = Information, Simil = Similarities, Arith = Arithmetic, Vocab = Vocabulary, Compr = Comprehension, PicC = Picture Completion, PicA = Picture Arrangement, BlkD = Block Design, ObjA = Object Assembly, Cod = Coding, VIQ = Verbal IQ, PIQ = Performance IQ, and Full = Full-Scale IQ.

between Verbal and Performance IQ.

On each subtest of the WISC-R, the mean for the general population is 10 ($SD = 3$). It can be seen from Table 2 that the best-outcome subjects scored highest on Similarities, Block Design, and Object Assembly. They scored lowest on Picture Arrangement and Arithmetic. Thus, the subjects consistently scored at or above average.

Adaptive and Maladaptive Behavior. Table 3 presents the data for the best-outcome group on the Vineland Adaptive Behavior Scales. It can be seen that the best-outcome group scored about average on the Composite Scale and on the subscales for Communication, Daily Living, and Socialization. However, Table 3 shows that some of the best-outcome subjects had marginal scores, including J. L., B. W., and M. M. Even so, all of the best-outcome subjects had Composite scores within the normal range.

As can be seen in Table 3, on the Maladaptive Behavior Scale (Parts I and II), the mean score for the best-outcome group indicated that, on average, these subjects did not display clinically significant levels of maladaptive behavior. Three of them scored in the clinically significant range versus one subject in the nonclinical comparison group, which had a mean of 7.7 on this scale.

Personality Functioning. The results of the Personality Inventory for Children are summarized in Table 4. The best-outcome subjects obtained valid profiles on the Per-

sonality Inventory for Children, as measured by the three validity scales (Lie, Frequency, and Defensiveness). As can be seen from the table, the subjects scored in the normal range across all scales. They tended to score highest on Intellectual-Screening, Psychosis, and Frequency. Intellectual-Screening assesses slow intellectual development, and Psychosis and Frequency assess unusual or strange behaviors. Only Intellectual-Screening was above the normal range, and this scale is affected by subjects' early history. For example, the scale contains statements such as "My child first talked before he (she) was two years old," which would be false for the best-outcome subjects regardless of their current level of functioning.

As Table 4 indicates, 4 best-outcome subjects had a single scale elevated beyond

Table 3
Scores on the Vineland Adaptive Behavior Scale for the Best-Outcome Subjects

Subject	Adaptive behavior				Maladaptive behavior
	Com	DLS	Soc	Comp	
R.S.	83	98	102	92	6
M.C.	119	93	86	98	16
M.M.	119	79	114	105	2
L.B.	107	108	112	108	4
J.L.	77	103	94	88	13
D.E.	93	81	82	80	15
A.G.	101	97	99	98	5
B.W.	83	74	105	83	9
B.R.	—	—	—	—	—
Mean	98	92	99	94	8.8

Note. Com = Communication, DLS = Daily Living Skills, Soc = Socialization, Comp = Adaptive Behavior Composite.

Table 4

T Scores on the Personality Inventory for Children for the Best-Outcome Subjects

Subject	T score																	
	Mean	<70	L	F	Def	Adj	Ach	I-S	Dvl	Som	Dep	Fam	DIq	Wdr	Anx	Psy	Hyp	Soc
R.S.	56	1	49	54	43	61	53	75	49	44	69	47	46	69	60	65	46	64
M.C.	52	1	48	63	37	43	39	54	38	64	55	54	46	65	51	75	40	55
M.M.	49	0	42	54	43	50	42	64	46	58	48	55	46	47	53	46	54	36
L.B.	51	1	60	50	49	49	37	70	39	55	49	48	51	45	60	51	49	51
J.L.	70	9	42	84	37	85	77	94	65	78	86	65	61	69	78	76	52	72
D.E.	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
A.G.	51	0	38	45	49	57	48	39	53	51	49	69	40	55	55	55	49	63
B.W.	54	1	45	63	50	59	64	48	55	47	44	57	90	44	45	46	62	44
B.R.	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Mean	55	2	46	56	44	58	51	64	49	57	57	56	54	56	57	59	50	55

Note. Mean = mean elevation across all scales. L = Lie scale, F = Frequency, Def = Defensiveness, Adj = Adjustment, Ach = Achievement, I-S = Intellectual-Screening, Dvl = Development, Som = Somatic Concern, Dep = Depression, Fam = Family Relations, DIq = Delinquency, Wdr = Withdrawal, Anx = Anxiety, Psy = Psychosis, Hyp = Hyperactivity, Soc = Social Skills.

the clinically significant range and a 5th (J. L.) had nine scales elevated, including the highest scores in the best-outcome group on Intellectual-Screening, Psychosis, and Frequency. Thus, this subject appeared to account for much of the elevation in scores on these scales. By comparison, there were 3 subjects in the nonclinical comparison group with at least one scale elevated.

Clinical Rating Scale. On this scale, 8 of the best-outcome subjects scored between 0 and 10, and the 9th (J. L.) scored 42. The mean was 8.8, with a standard deviation of 12.9. The nonclinical comparison subjects all scored between 0 and 5 (mean = 1.7, $SD = 2.1$). Because these SD s are unequal, we used a nonparametric statistic, a Mann-Whitney U test, revealing a significant difference between groups, $U = 19$, $p < .05$. Thus, the best-outcome subjects displayed more deviance than did the comparison subjects, but most of the deviance appeared to come from one subject, J. L.

Discussion

This study is a later and more extensive follow-up of two groups of young subjects with autism who were previously studied by Lovaas (1987): (a) an experimental group ($n = 19$) that had received very intensive behavioral treatment and (b) a control group ($n = 19$) that had received minimal behavioral

treatment. In the present study we have reported data on these children at a mean age of 13 years for subjects in the experimental group and 10 years for those in the control group. The data were obtained from a comprehensive assessment battery.

The main findings from the test battery were as follows: First, subjects in the experimental group had maintained their level of intellectual functioning between their previous assessment at age 7 and the present evaluation at a mean age of 13, as measured by standardized intelligence tests. Their mean IQ was about 30 points higher than that of control subjects. Second, experimental subjects also displayed significantly higher levels of functioning than did control subjects on measures of adaptive behavior and personality. Third, in a particularly rigorous evaluation of the 9 subjects in the experimental group who had been classified as best-outcome (normal-functioning) in the earlier study (Lovaas, 1987), the test results consistently indicated that the subjects exhibited average intelligence and average levels of adaptive functioning. Some deviance from average was found on the personality test and the clinical ratings. However, this deviance appeared to derive from the extreme scores of one subject, J. L. (see Table 2, 3, and 4). This subject also had been removed from nonspecial education classes and placed in a class for children with language delays, and he obtained relatively

low scores (about 80) on the Verbal section of the intelligence test and the Communication section of the measure of adaptive behavior. Thus, he no longer appeared to be normal-functioning. However, the remaining 8 subjects who had previously been classified as normal-functioning demonstrated average IQ, with intellectual performance evenly distributed across subtests, were able to hold their own in regular classes, did not show signs of emotional disturbance, and demonstrated adequate development of adaptive and social skills within the normal range. In addition, subjective clinical impressions of blind examiners did not discriminate them from children with no history of behavioral disturbance. These 8 subjects (42% of the experimental group) may be judged to have made major and enduring gains and may be described as "normal-functioning." By contrast, none of the control group subjects achieved such a favorable outcome, consistent with the poor prognosis for children with autism reported by other investigators (Freeman, Ritvo, Needleman, & Yokota, 1985).

In order to evaluate this outcome, we must pay close attention to whether or not our methodology was sound. The adequacy of our methodology is crucial because the outcome in the present study represents a major improvement over outcomes obtained in previous experimental studies on the treatment of children with autism (Rutter, 1985). The only reports of comparable outcomes have come from uncontrolled case studies (e.g., Bettelheim, 1967), and subsequent investigations have indicated that these case studies grossly overestimated the outcomes obtainable with the treatment that was provided. Similarly, reports of major gains in other populations, such as large IQ increases in children from impoverished backgrounds, also have been based on highly questionable evidence (Kamin, 1974; Spitz, 1986). Such reports have the potential to cause a great deal of harm by misleading consumers and professionals.

A detailed description of all the methodological safeguards that should be built

into a treatment study is beyond the scope of the present report (see Kazdin, 1980; Kenner & Norton-Ford, 1982; Spitz, 1986). However, we note that we incorporated a large number of methodological safeguards in both the original study (Lovaas, 1987) and the present investigation:

1. The experimental group and control group received equivalent assessment batteries at intake and were found to be very similar on a multitude of important variables. Moreover, the number of control group subjects who were predicted to achieve normal functioning, had they received intensive treatment, was approximately equal to the number of experimental subjects who actually did achieve normal functioning with intensive treatment (Lovaas & Smith, 1988). Thus, the subject assignment procedure yielded groups that were comparable prior to treatment. This provided a strong indication that the superior functioning of the experimental group after treatment was a result of the treatment itself rather than a biased procedure for assigning subjects to the experimental group.

2. All subjects remained in the groups to which they were assigned at intake. Only one subject dropped out, and they were replaced. Therefore, the original composition of the groups was essentially preserved.

3. All subjects were independently diagnosed as autistic by PhD or MD clinicians, and there was high agreement on the diagnosis between the independent clinicians. This provided evidence that subjects met criteria for a diagnosis of autism.

4. Prior to treatment, these subjects appeared to be comparable to those diagnosed as having autism in other research investigations. Evidence for this comes from the second control group that was incorporated into the initial treatment study. This group was evaluated by another research team (independent of ours), had similar IQ at intake based on the same measures of intelligence that we used, yet showed similar outcome data to those reported by other investigators. Additional evidence can be

derived from the similarity of our intake data to data reported by other investigators (Lovaas et al., 1989). For example, although Schopler and his associates (Schopler, Short, & Mesibov, 1989) suggested that our sample had a higher mean IQ than did other samples of children with autism, their own data do not appear to differ from ours (Lord & Schopler, 1989). Thus, there is evidence that our subjects were a typical group of preschool-age children with autism rather than a select group of high-level children with autism who would have been expected to achieve normal functioning with little or no treatment.

5. The first control group, which received up to 10 hours a week of one-to-one behavioral treatment, did not differ at post-treatment from the second control group, which received no treatment from us. Both groups achieved substantially less favorable outcomes than did the experimental group. Because all groups were similar at pretreatment, this result confirms that our subjects had problems that responded only to intensive treatment rather than problems such as being noncompliant or holding back (masking an underlying, essentially average intellectual functioning that would respond to smaller-scale interventions).

6. Subjects' families ranged from high to low socioeconomic status, and, on average, they did not differ from the general population (Lovaas, 1987). Thus, although our treatment required extensive family participation, a diverse group of families was apparently able to meet this requirement.

7. The treatment has been described in detail (Lovaas et al., 1980; Lovaas & Leaf, 1981), and the effectiveness of many components of the treatment has been demonstrated experimentally by a large number of investigators over the past 30 years (cf. Newsom & Rincover, 1989). Hence, our treatment may be replicable, a point that is discussed in greater detail later.

8. The results of the present follow-up, which extended several years beyond discharge from treatment for most subjects, are an encouraging sign that treatment gains

have been maintained for an extended period of time.

9. A wide range of measures was administered, avoiding overreliance on intelligence tests, which have limitations if used in isolation (e.g., bias resulting from teaching to the test, selecting a test that would yield especially favorable results, failing to assess other aspects of functioning such as social competence or school performance) (Spitz, 1986; Zigler & Trickett, 1978).

10. The use at follow-up of a normal comparison group, standardized testing, and blind rating allowed for an objective, detailed, and quantifiable assessment of treatment effectiveness. A particularly rigorous assessment was given to those subjects who showed the most improvement.

Taken together, these safeguards provide considerable assurance that the favorable outcome of the experimental subjects can be attributed to the treatment they received rather than to extraneous factors such as improvement that would have occurred regardless of treatment, biased procedures for selecting subjects or assigning them to groups, or narrow or inappropriate assessment batteries.

Despite the numerous precautions that we have taken, several concerns may be raised about the validity of the results. Perhaps the most important is that the assignment to the experimental or control group was made on the basis of therapist availability rather than a more arbitrary procedure such as alternating referrals (assigning the first referral to the experimental group, the second to the control group, the third to the experimental group, and so forth). However, it seems unlikely that the assignment was biased in view of the pretreatment data we have presented on the similarity between the experimental and control groups. On the other hand, we do not know as yet whether there exists a pretreatment variable that does predict outcome but was not among the 19 we chose, yet could have discriminated between groups. In an earlier publication (Lovaas et al., 1989), we responded in some

detail to the concern about subject assignment as well as other possible problems associated with the original study. There are certain additional questions that may be raised by this follow-up investigation:

1. The experimental group was older than the control group at the time of this follow-up evaluation. We explained this finding earlier and noted that data analyses indicated that it was unlikely that this age difference reflected a bias in subject assignments.

2. The follow-up assessments for 17 of the lower functioning subjects in this study were conducted by staff members from our Project, who could have biased the test results. However, as noted previously, a check revealed no evidence of such a bias.

3. The Clinical Rating Scale, based on an interview with subjects who had been classified as normal-functioning in the original study, has no norms or data on reliability and validity. However, we regard the interview simply as an extra check on whether the examiners detected residual signs of autism or other behavior problems that were somehow overlooked in the three other (well-standardized) measures in the study and their 30 subscales. We do not regard the interview as an instrument that by itself yields conclusive results. No other interview that suited our purposes currently exists. In future investigations, we plan to use an interview that Michael Rutter and his associates are now developing for the purpose of detecting of residual signs of autism in individuals with average intelligence.

4. As in most long-term follow-up studies, we had some missing data. However, there is no evidence that the missing data would have changed the overall results.

5. In our analysis of the best-outcome group, we noted that the group averages deviated from "normal" on one subscale of the Personality Inventory for Children and on the Clinical Rating Scale. We then attributed this deviance to the extreme scores of one subject rather than to general problems within this group. We recognize that group

averages are seldom interpreted this way. However, as statisticians and methodologists have pointed out (e.g., Barlow & Hersen, 1984), there are many times when group averages represent the performance of few or no subjects within the group. This was one of those times, as is clearly shown by the data on individual subjects (Tables 2, 3, and 4). Deviance was found almost exclusively in one subject, not evenly distributed across all subjects, and we have presented the results accordingly.

The most important void for research to fill at this time is replication by independent investigators who employ sound methodologies. Given the objective assessment instruments that we used and the detailed description that we have provided of the treatment (Lovaas et al., 1980), such a replication should be possible. However, the treatment is complex and to replicate it properly, an investigator probably needs to possess (a) a strong foundation in learning theory research; (b) a detailed knowledge of the treatment manual we used; (c) a supervised practicum of at least 6 months in one-to-one work with clients who have developmental delays, emphasizing discrimination learning and building complex language; and (d) a commitment to provide 40 hours of one-to-one treatment to client per week, 50 weeks per year, for at least 2 years. Our best-outcome subjects all required a minimum of 2 years of intensive treatment to achieve average levels of functioning (another indication that those subjects had pervasive disabilities and were not merely non-compliant).

A second void to fill concerns the majority of children who did not benefit to the point of achieving normal functioning with intensive treatment. Perhaps an earlier start in treatment would have been all that was needed to obtain favorable outcomes with many of these children. More pessimistically, perhaps such children require new and different interventions that have yet to be discovered and implemented. In any case, it is essential to develop more appropriate

services for these children.

Finally, a rather speculative but promising area for research is to determine the extent to which early intervention alters neurological structures in young children with autism. Autism is almost certainly the result of deficits in such neurological structures (Rutter & Schopler, 1987). However, laboratory studies on animals have shown that alterations in neurological structure are quite possible as a result of changes in the environment in the first years of life (Sirevaag & Greenough, 1988), and there is reason to believe that alterations are also possible in young children. For example, children under 3 years of age overproduce neurons, dendrites, axons, and synapses. Huttenlocher (1984) hypothesized that, with appropriate stimulation from the environment, this overproduction might allow infants and preschoolers to compensate for neurological anomalies much more completely than do older children. Caution is needed in generalizing from these findings on average children to early intervention with children with autism, particularly because the exact nature of the neurological anomalies of children with autism is unclear at present (e.g., Rutter & Schopler, 1987). Nevertheless, the findings suggest that intensive early intervention could compensate for neurological anomalies in such children. Finding evidence for such compensation would help explain why the treatment in this study was effective. More generally, it might contribute to an understanding of brain-behavior relations in young children.

References

- Barlow, D. H., & Hersen, M. (1984). *Single case experimental design: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.
- Bettelheim, B. (1967). *The empty fortress*. New York: The Free Press.
- DeMyer, M. K., Hingtgen, J. N., & Jackson, R. K. (1981). Infantile autism reviewed: A decade of research. *Schizophrenia Bulletin*, 7, 388-451.
- Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised*. Circle River, 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 5-year study. *Journal of the American Academy of Child Psychiatry*, 24, 290-311.
- Huttenlocher, P. R. (1984). Synapse elimination and plasticity in developing human cerebral cortex. *American Journal of Mental Deficiency*, 88, 488-496.
- Kamin, L. J. (1974). *The science and politics of I.Q.* New York: Wiley.
- Kanner, L. (1971). Follow-up study of 11 autistic children originally reported in 1943. *Journal of Autism and Childhood Schizophrenia*, 1, 119-145.
- Kazdin, A. (1980). *Research design in clinical psychology*. New York: Harper & Row.
- Kendall, P. C., & Norton-Ford, J. D. (1982). Therapy outcome research methods. In P. C. Kendall & J. N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 429-460). New York: Wiley.
- Leiter, R. G. (1959). Part I of the manual for the 1948 revision of the Leiter International Performance Scale: Evidence of the reliability and validity of the Leiter tests. *Psychology Service Center Journal*, 11, 1-72.
- Lord, C., & Schopler, E. (1989). The role of age at assessment, developmental level, and test in the stability of intelligence scores in young autistic children. *Journal of Autism and Developmental Disorders*, 19, 483-499.
- Lotter, V. (1978). Follow-up studies. In M. Rutter & E. Schopler (Eds.), *Autism: A reappraisal of concepts and treatment*. London: Plenum Press.
- Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology*, 55, 3-9.
- Lovaas, O. I., Ackerman, A. B., Alexander, D., Firestone, P., Perkins, J., & Young, D. (1980). *Teaching developmentally disabled children: The me book*. Austin, TX: Pro-Ed.
- Lovaas, O. I., Koegel, R. L., Simmons, J. Q., & Long, J. S. (1973). Some generalization and follow-up measures on autistic children in behavior therapy. *Journal of Applied Behavior Analysis*, 6, 131-166.
- Lovaas, O. I., & Leaf, R. L. (1981). *Five video*

- tapes for teaching developmentally disabled children. Baltimore: University Park Press.
- Lovaas, O. I., & Smith, T. (1988). Intensive behavioral treatment with young autistic children. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 11, pp. 285-324). New York: Plenum Press.
- Lovaas, O. I., Smith, T., & McEachin, J. J. (1989). Clarifying comments on the young autism study: Reply to Schopler, Short and Mesibov. *Journal of Consulting and Clinical Psychology*, 57, 165-167.
- McEachin, J. J. (1987). *Outcome of autistic children receiving intensive behavioral treatment: Psychological status 3 to 12 years later*. Unpublished doctoral dissertation, University of California, Los Angeles.
- Newsom, C., & Rincover, A. (1989). Autism. In E. J. Mash & R. A. Barkley (Eds.), *Treatment of childhood disorders* (pp. 286-346). New York: Guilford Press.
- Rutter, M. (1970). Autistic children: Infancy to adulthood. *Seminars in Psychiatry*, 2, 435-450.
- Rutter, M. (1985). The treatment of autistic children. *Journal of Child Psychology & Psychiatry*, 26, 193-214.
- Rutter, M., & Schopler, E. (1987). Autism and pervasive developmental disorders: Concepts and diagnostic issues. *Journal of Autism and Developmental Disorders*, 17, 159-186.
- Schopler, E., Short, A., & Mesibov, G. (1989). Relation of behavioral treatment to "normal functioning": Comment on Lovaas. *Journal of Consulting and Clinical Psychology*, 57, 162-164.
- Short, A., & Marcus, L. (1986). Psychoeducational evaluation of autistic children and adolescents. In S. S. Strichart & P. Lazarus (Eds.), *Psychoeducational evaluation of school-aged children with low-incidence disorders* (pp. 155-180). Orlando, FL: Grune & Stratton.
- Simeonsson, R. J., Olley, J. G., & Rosenthal, S. L. (1987). Early intervention for children with autism. In M. J. Guralnick & F. C. Bennett (Eds.), *The effectiveness of early intervention for at-risk and handicapped children* (pp. 275-296). Orlando, FL: Academic Press.
- Sirevaag, A. M., & Greenough, W. T. (1988). A multivariate statistical summary of synaptic plasticity measures in rats exposed to complex, social and individual environments. *Brain Research*, 441, 386-392.
- Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Interview Edition Survey Form Manual*. Circle Pines, MN: American Guidance Service.
- Spitz, H. H. (1986). *The raising of intelligence*. Hillsdale, NJ: Erlbaum.
- 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). *Manual for the Wechsler Intelligence Scale for Children-Revised*. New York: Psychological Corp.
- Wirt, R. D., Lachar, D., Klinedinst, J. K., & Seat, P. D. (1977). *Multidimensional descriptions of child personality: A manual for the Personality Inventory for Children*. Los Angeles: Western Psychological Services.
- Zigler, E., & Trickett, P. K. (1978). IQ, social competence, and evaluation of early childhood intervention programs. *American Psychologist*, 33, 789-798.

Received: 5/15/91; first decision: 10/16/91; accepted: 1/23/92.