DOCUMENT RESUME

ED 170 990 EC 114 011

AUTHOR Macy, Daniel J.
TITLE Application of a Theoretical Control Strategy in Early Intervention for the Handicapped.
INSTITUTION Dallas Independent School District, Tex.
SPONS AGENCY Bureau of Education for the Handicapped (DHEW/OE), Washington, D.C.
PUB DATE Apr 79
NOTE 42p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, California, April, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Developmental Disabilities; *Developmental Programs; *Early Childhood Education; Exceptional Child Research; *Intervention; *Physically Handicapped; *Prediction; Program Effectiveness
IDENTIFIERS Project Kindling Individual Development Systems

ABSTRACT

The effectiveness of the Dallas, Texas Project KIDS (Kindling Individual Development Systems) on the developmental progress of 17 developmentally delayed and physically handicapped children (aged 18 months or less at pretest) was studied, and the accuracy of projections of children's progress made by professionals was investigated. The Bayley Scales of Infant Development, the KIDS Inventory of Development, and case study data were used to score projections. Among findings were that projected scores were reliable and relatively free of measurement error; that sampled children made highly significant pre-post improvement; that there was variability in scores projected by experts for any given child, and that the extent of projected score variability differed across children. (SBH)

Application of a Theoretical Control Strategy in Early Intervention for the Handicapped

by

Daniel J. Macy, Ph.D.
Principal Evaluator - Special Education
Department of Research, Evaluation, and Information Systems
Dallas Independent School District
Dallas, Texas

Paper presented at a roundtable session in the annual meeting of the American Educational Research Association, San Francisco, 1979.

Support for this study was partially provided by funds from the Bureau of Education for the Handicapped, U. S. Office of Education. The opinions expressed herein do not necessarily reflect those of the U. S. Office of Education, and no official endorsement should be inferred.

Introduction

Project KIDS (Kindling Individual Development Systems), under the direction of Dr. Ruth M. Turner, is a Dallas Independent School District model program for handicapped infants, toddlers, preschool children and their families. Beginning in the 1975-76 year, The Bureau of Education for the Handicapped (BEH) provided three-year support for the project under the Handicapped Children's Early Intervention program, and the Dallas ISD assumed support for the project beginning the 1978-79 school year. However, BEH, provided an additional three-year grant to fund Project KIDS OUTREACH, which is an extension of the KIDS model into additional Dallas ISD early childhood classrooms and into surrounding suburban school districts.

Dr. Turner designed the KIDS model to serve developmentally delayed and physically handicapped preschool children age birth to six years. Services are delivered through home-based instruction, center-based infant stimulation classes, and school-based early childhood class units. The primary instructional vehicle is the Mini Activity Plan (MAP) which details the individualized educational plan for each child in reference to observed performance in the KIDS Inventory of Development. Each developmental item in the Inventory cross references a variety of curricula and activities to promote the developmental skill or behavior tested by that particular item. Hence, project staff can plan instructional intervention in direct response to assessed performance levels.

Another key feature of the KIDS model is the integration of the child's parents and family into a cooperative instructional role with project staff. In the home-based component, the parent is seen as the primary instructional agent with more instruction being assumed by staff as the child progresses toward school-based instruction. A competency based parent involvement program has been developed to provide individualized training for parents (Turner, 1978). This program provides a continuum of training activities which parents select in terms of self-reported competency levels (Carter, 1978; Macy, 1978).

Project KIDS also receives supplementary appraisal services from the University Affiliated Center of the University of Texas Health Science Center and faculty from the Special Education Program at the University of Texas at Dallas also provides consultative assistance to the project.

The purpose of the current study was to secure comparative data for evaluating the impact of project intervention on developmental progress of children. As one might expect, it is virtually impossible to identify a random control group for the population served in Project KIDS, and identification of even a comparison group is almost impossible, or certainly unrealistic. In the absence of a control group or a comparison group of children, one must find an alternative point of reference for comparison. One approach to obtaining a comparative point of reference in early intervention has been to calculate an expectancy score based on pretest performance. The basic procedure is to compute a developmental rate as the ratio of developmental age to chronological age at the time of pretest. Simeonsson and Wiegerink (1975) built on this concept in suggesting an index of efficiency for comparing heterogeneous children within a project or across projects.

Another approach to obtaining a comparison point of reference is to adopt a time lag design and form a comparison group or groups within the project as a function of age at time of pre-post measurement. The basic premise of this time lag approach is that the pretest scores of one group can serve as the comparison for posttest scores of another group. For a discussion of time lag design, see Goulet (1975).

Another strategy frequently used in special education research is the matched pairs design. The popularity of this design is largely due to the difficulty in securing adequate control groups in special education populations.

While the above research strategies have merit and potential application, none is the design of choice due to problems encountered in special subject populations such as the population served in Project KIDS. The comparison of developmental rates (developmental age ÷ chronological age) at pretest and posttest is one of the most popular research strategies in use with the early childhood handicapped population. However, two reservations come to mind when approach is used with very young children. The first centers about the stability in the rate of maturation during the early months and years of life. If in fact one can assume that the maturation rate is linear throughout the first 60 months of life, then one might place greater confidence in comparison of pre-post developmental rates, but the assumption of linearity in maturational rate appears tenuous at best.

Comparison of pre-post developmental rates certainly has the potential to demonstrate significant pre-post gains, but an alternative to project intervention as an explanation of these gains could easily be an artifact of possible curvilinearity in maturational rate. For example, if an appreciable number of children in the project were pretested at an age just prior to a normal increase in maturational rate in the dependent variable, the observed pre-post gains could be largely due to just normal maturation and not necessarily to project intervention. This phenomenon could also work to decrease or wash out real gains of children due to project intervention.

A second reservation about the developmental rate strategy has to do with the impact of measurement error on the ratio of developmental age to chronological age, and especially so when ages involve small numbers. For example, the standard error of measurement for the Bayley mental scale is 6.7 raw score units for a 12-month old child (Bayley, 1969). If such a child were to have a pretest developmental age of 7 months, the developmental rate would be .58, but if the pretest raw score were only one standard error higher, the child's developmental age would be 9 months. This would change the pretest developmental rate from .58 to .75, simply as a result of measurement error. Such variation could influence the interpretation of pre-post developmental rate comparisons.

Another approach to obtaining a comparison point of reference would be to adopt a kind of time lag design and form a comparison group or groups within the project as a function of age at time of pre-post measurement. The following outlines a schema of a possible time lag approach in Project KIDS.

	Age in months			
Treatment Population	12	24	36	48
Home-based	pre ⟶	post		
Center-based		pre ⟶	post	
School-based			pre ⟶	post

In the time lag strategy the pretest scores of one group serve as the comparison point for the posttest scores of another group. In Project KIDS, the comparison of pretest scores of center-based children to posttest scores of home-based children could be used to test the effectiveness of home-based instruction. Similarly, one could compare pretest school-based to posttest center-based to test effectives of center-based instruction, but there would be no comparison for school-based instruction. One obvious limitation to the time lag approach is the inability to assign randomly children to pre and post comparison groups, thereby reducing experimental control. Another limitation is that the time lag strategy requires a sufficient number of children in each comparison group to yield suitable sample sizes in the lagged pre-post comparisons. In Project KIDS children enroll on a referral basis which reduces the probability of achieving suitable sample sizes for the necessary comparisons.

The matched pairs design, a third popular strategy, can provide a reasonably defensible design, if children can be closely matched on all critical variables. Macy and Carter (1975) have reported success in matching severely handicapped school age children, but good success in matching requires a large population from which to draw. In Project KIDS the population of children is quite limited in size, and the heterogeneity among children and families would make the matching process most difficult, or perhaps not even possible.

In light of the limitations of the research designs usually applied in special education research, it was thought a more innovative strategy was needed. The concept of a theoretical control strategy was consequently developed for the product evaluation of Project KIDS. The theoretical control strategy requires a panel of experts to project the test score performance of subjects given the assumption of no experimental intervention. In other words, experts are asked to score a given test as they think the subject would have scored at some point in time given no intervention.

The theoretical control score is then the average score (or a desired central tendency measure) of the scores projected by the expert panel.

Panels of experts are commonly used in instrument validation studies, but there has been no or very limited use of experts in developing theoretical controls. Curtis and Donlon (1972) have studied agreement among experts in rating video tape recordings of multiply handicapped children. Stricklin (1974) used an expert panel to develop psychiatric profiles of adults in a clinical setting. Reported research suggested that the theoretical control strategy as conceived in the Project KIDS evaluation might be feasible, but a review of literature revealed no applications of an expert panel in obtaining projected performance in the context of a control comparison.

The theoretical control approach seemed to be especially applicable to Project KIDS in that the project addresses developmental delay, an area with sufficient background research and clinical experience to make projections possible. The procedure was to select professionals who were recognized and accepted as experts in child development. These experts were independent of Project KIDS and possessed substantial credentials and experience in child development. Each expert independently reviewed the complete assessment records of each child and then formulated an assessment profile of the child projected forward twelve months from date of pretest assessment. Experts based their projections on the assumption of no intervention.

The current application of the theoretical control group strategy in Project KIDS involved a sample of 17 children (aged 18 months or less at pretest) and an expert panel of four nationally recognized leaders in child development. Panel members used the Bayley Scales of Infant Development and the KIDS Inventory of Development (a developmental checklist) to record their 12-month projections. In effect, experts scored the Bayley and KIDS Inventory as though they were actually administering a posttest to the child.

The study attempted to answer the following research questions:

1. What was the psychometric status of scores projected by experts?

   1.1 What was the reliability of projected scores?

   1.2 What was the validity of projected scores?

   1.3 Did individual experts exhibit any consistent deviation in projecting scores?

   1.4 Was the variability in projected scores related to children's pretest performance, related subtest performance, age, number of items projected, handicap, or extent of available information?

2. Was children's developmental progress, if any, due to project intervention?

   2.1 Was there significant improvement in actual pre-post scores of sampled children?

   2.2 What were the characteristics of theoretical control scores?

   2.3 How did theoretical control scores compare to actual posttest scores?

Procedures

The following describes instrumentation, sample, and data collection procedures used in the study.

Instrumentation

Expert panel members used the Bayley Scales of Infant Development and the KIDS Inventory of Development to score their projections. Other "instrumentation" included the child case study data given to experts upon which to base their projections.

Bayley Scales of Infant Development. The Bayley is one of the best known instruments for measuring developmental status in the first two and one-half years of life. The instrument yields a Mental Development Index, a Psychomotor Development Index, and a Behavior Record (Bayley, 1969). The standardization sample included 1,262 children and was representative of the United States population (1960 census) in terms of urban-rural residence, white-nonwhite race, occupation and education of the head of the household, and geographic region (of course, the standardization sample included only normal children). Split-half reliability coefficients reported for the mental scale range from .81 to .93, and coefficients for the motor scale range from .68 to .92. The mental scale contains 163 items and the motor scale has 81 items.

KIDS Inventory of Development. The KIDS Inventory was developed primarily under the direction of Dr. Nanci Bray, a University of Texas at Dallas consultant to Project KIDS, and Dr. Ruth Turner, Project KIDS Director. Project KIDS staff also provided valuable contribution and effort to developing the Inventory. The Inventory is still in a developmental status and is undergoing further refinement and modification. The basic format of the Inventory is a checklist of developmental behaviors sequenced according to chronological age 0 to 72 months in four areas of development: gross motor (77 items), fine motor (70 items), language/cognitive (112 items), and self-help (64 items). Scoring of the inventory is pass-fail for each behavior tested. Preliminary study shows that the criterion validity of the KIDS Inventory with the Bayley mental is .75 and .88 with the Bayley motor (Carter, 1978).

Child Case Study Data. An extensive case study profile was compiled for each child in the study. This profile included results from educational, psychological, sociological, and medical assessments. Pretest record forms from the Bayley mental and motor scales were also included in the profiles. A small pilot theoretical control study involving a panel of seven professionals and five children determined that the above information was relevant and useful in preparing experts for the task of projecting performance. Since the extent of assessment and evaluation completed on each child varied, the amount of information contained in each child's profile also varied. The number of pages in a profile ranged from 40 to 100, and the typical number of pages was 60.

Sample

The sample for the study included 17 children and four expert panel members. The four panel members were nationally recognized experts in early childhood education for the handicapped. They were:

Dr. Bettye Caldwell
Center for Child Development and Education
University of Arkansas at Little Rock
Little Rock, Arkansas

Dr. Ernest Gotts
School of Human Development
University of Texas at Dallas
Richardson, Texas

Dr. Maynard Reynolds
Department of Psychoeducational Studies
University of Minnesota
Minneapolis, Minnesota

Dr. Francis Walker
Office of Child Development
California State Department of Education
Sacramento, California

Experts received a stipend for their participation in the study.

The sample of children included 17 children (seven male and ten female) 18 months and younger at time of Bayley pretest (i.e., entry into Project KIDS). The average age at pretest was 11.68 months, and the standard deviation was 5.44 months. The average pretest Bayley Mental and Motor raw scores were 66.29 and 25.35; the standard deviations were 33.38 and 14.64, respectively. The average mental developmental age was 5.4 months, and the average motor developmental age was 5.6 months. Since the average chronological age was 11.68 months, the average mental maturational rate was 46 percent (5.4 ÷ 11.68), and the average motor rate was 48 percent (5.6 ÷ 11.68). No scores from the KIDS Inventory were available for this sample, since the Inventory was undergoing initial development at this time.

There was a wide range of handicapping conditions evident in the sample. The following lists a brief description of the conditions for each child:

Child	Handicapping conditions[1]
1	chromosomal anomaly, minor physical abnormalities with possible hearing problem
2	language and personal/social developmental delay
3	severe retardation with vision and hearing problems
4	hydrocephalic, CNS damage, trainable level
5	seizure disorder, left hemiparesis
6	CP, mild spastic paraplegia
7	infantile myoclonus, profound brain damage
8	multiple congenital anomalies of unknown etiology
9	Down's snydrome (MR)

10	CNS damage, hypotonia
11	microcephalic
12	language and motor delay
13	CNS damage, physical anomalies
14	CP, right spastic hemiparesis
15	Down's snydrome (MR)
16	Down's snydrome (MR), failure to thrive
17	developmental delay of unknown etiology

[1] CNS - central nervous system, CP - cerebral palsy, MR - mental retardation

Inspection of the above list of conditions indicates at least superficially, the extent of handicap involvement in sampled children. The vast majority of the sample obviously included children with severe handicapping conditions. However, the extent of severity in the sample was likely greater than that in the total Project KIDS population, since initial referrals to the project tended to have more severe involvement.

Data Collection

Expert panel members received detailed directions for scoring their projections and all necessary materials. Instructions directed experts to project what they thought the child's performance on the Bayley and KIDS Inventory would be 12 months from the time of pretesting. Experts were to assume no special interventions into the child and family life (aside from any necessary life support for the child). A complete set of directions to experts appears in the Appendix.

Results from a preliminary pilot study with local experts showed that one hour was the typical time required to review a child's case study and complete projected test score performance. However, expert panel members were not asked to record their time in completing the score projections. Two of the four experts were prompt in returning their completed projections, and the other two experts delayed in completing their projections.

Results

The results of the study are reported by individual research questions.

1.1 What was the realiability of projected scores?

Data analysis for this question utilized analysis of variance to estimate reliability of expert projections. The procedure followed was that as described by Winer (197, pp. 283 ff.). Application of this procedure yielded what is known as an intraclass correlation coefficient, which provided an estimate of reliability for any one expert. The Spearman-Brown prediction formula was used to obtain the reliability estimate for the average projected score from all four experts.

The general computational procedure was to calculate the unbiased estimate of the ratio of true score variance to error score variance. This ratio, termed theta, was then used in the following equation to compute reliability coefficients:

$r_k = \frac{k \emptyset}{1 + k\emptyset}$, where k is the number of experts, an $\emptyset$ is the ratio of true score variance to error score variance. Summary tables for ANOVAs and detailed reliability calculations are reported in the Appendix.

In general, reliability estimates were quite high, and results showed that the extent of consistency among scores projected by experts was easily within acceptable limits. In short, projected scores were reliable and relatively free of measurement error (estimated by ANOVA). Table 1 reports reliability estimates for the Bayley and KIDS Inventory.

Table 1

Estimated Reliability for Projected Scores

Test	Reliability Coefficient	
	One Expert	Average of Four Experts
Bayley Mental	.51	.81
Bayley Motor	.84	.95
KIDS Cognitive/Language	.73	.91
KIDS Gross Motor	.69	.90
KIDS Fine Motor	.77	.93
KIDS Self Care	.70	.87[a]
Average	.71	.90

[a] Coefficient based on three rather than four experts due to incomplete data from one expert.

Inspection of Table 1 shows that reliabilities for any one expert ranged from .51 to .84, and they ranged from .81 to .95 for the average of four experts. The overall average reliability coefficient for projected scores from one expert was .71 and was .90 for the average of four experts. The above results demonstrate the feasibility of the theoretical control strategy in that projected scores were internally consistent among experts, and the proportion of measurement error in projected scores was well within acceptable limits.

1.2 What was the validity of projected scores?

The best measure of validity of course would be actual measurement of children taken 12 months from date of pretest without project intervention. Unfortunately, it was impossible to collect such measurements. Another measure of validity would be to compare projections of the current panel with an additional panel of experts, but financial resources did not permit such a comparison at the time of this study.

At the present time, the major source of validity rests upon the credentials of selected expert panel members. Inspection of the list of panel members on page five speaks to the high level of expertise and professional status among experts selected for the study. While the validity of the Bayley Scales has been well established and while criterion validity of the KIDS Inventory has also been established, the validity of scores projected on the Bayley and Inventory rests primarily upon professional credential and experience embodied in the expert panel. No descriptive data about recent experience with the Bayley or with young handicapped children were collected from panel members.

1.3 Did individual experts exhibit any consistent deviation in projecting scores?

The above question essentially asks if one or more experts tended to be consistently higher or lower than other experts in projecting scores. A repeated-measures ANOVA was used to test for a significant expert main effect in projected scores (this was the same ANOVA discussed under question two). Results of the analysis revealed no significant expert effects, with the possible exception of scores from the KIDS self-care scale where the F-ratio was significant at the 10 percent level, but only a five percent or smaller level is generally considered significant.

Table 2 presents the means and standard deviations of projected scores as well as the F-ratio for testing significance of difference among average scores by experts. The ANOVA summary tables are included in the Appendix.

The results clearly indicate that individual experts did not have any significant bias in projecting scores. In other words, there was no significant tendency for any given expert to project consistently higher or lower than other experts. The data in Table 2 also show that variability in projected scores within each expert was comparable across all four experts. An F_{max} test for homogeneity of variance in projected scores for each scale verified the above observation.

1.4 Was the variability in projected scores related to children's pretest performance, related subtest performance, age, number of items rated, handicap, or extent of available information?

As expected, there was variability in projected scores for any one child, and Table 3 reports the standard deviations among projected scores for each child for both the Bayley and KIDS Inventory. These data (Table 3) indicate that the greatest extent of projected score variability was in the Bayley mental scale, and the least variability was in the KIDS self-care scale.

While the interrater reliability estimates (see question 1.1) showed that the degree of variability was within tolerable limits, questions of interest pertained to identifying any variables or factors which might have been associated with score variability among experts. For example, were

Table 2

Means, Standard Deviations, and F-ratios for Projected Scores[a]

				Expert			
Test		(1)	(2)	(3)	(4)	F	p
Bayley mental	M	98.9	100.8	100.5	96.4	0.14	NS
	SD	34.2	31.9	33.1	30.6		
Bayley motor	M	41.5	41.9	40.9	38.8	1.00	NS
	SD	14.4	15.1	13.4	15.3		
KIDS cognitive/ language	M	33.4	34.7	35.6	31.8	1.45	NS
	SD	7.6	11.0	13.1	12.2		
KIDS gross motor	M	34.1	31.1	29.9	28.8	1.62	NS
	SD	10.3	14.2	13.6	15.9		
KIDS fine motor	M	28.1	27.0	28.2	26.4	0.40	NS
	SD	12.5	12.3	10.6	10.9		
KIDS self care	M	16.9	—[b]	20.5	18.8	2.94	<.10
	SD	6.2	—	10.5	8.4		

[a] M = mean or average score, SD = standard deviation, NS = not significant.

[b] Incomplete data from expert number two.

Table 3

Standard Deviations Among Projected
Scores for Each Child

	Bayley		KIDS Inventory			
Child	Mental	Motor	Gross Motor	Fine Motor	Cognitive/ Language	Self-Care
1	5.32	4.57	10.98	4.08	6.18	2.38
2	5.00	5.89	2.52	3.79	8.96	2.06
3	3.11	2.36	7.51	4.79	1.73	1.53
4	14.48	6.61	7.55	8.66	9.07	3.51
5	4.51	8.42	6.70	2.99	2.45	4.62
6	7.97	1.15	2.94	2.94	5.94	8.89
7	7.63	1.50	2.65	3.32	3.59	3.30
8	3.20	3.77	2.06	1.91	5.00	4.04
9	9.15	3.10	6.70	1.53	5.20	2.08
10	7.63	9.57	9.68	10.37	4.57	4.03
11	11.92	10.60	11.18	9.67	7.59	3.87
12	6.40	3.00	2.22	4.99	2.99	4.55
13	7.93	1.63	6.63	3.00	0.58	4.93
14	7.68	4.19	8.54	7.05	6.95	9.14
15	7.50	9.57	8.54	3.92	3.40	2.65
16	9.33	3.00	2.83	10.99	3.70	3.59
17	8.04	4.24	5.03	1.73	5.25	5.26
Average	7.46	4.89	6.13	5.04	4.87	4.14
SD	2.89	3.04	3.14	3.11	2.38	2.12

experts more consistent in projecting scores for older or younger children? Did the type of handicap involvement affect expert consistency in projecting scores?

The basic research procedure for addressing the above questions consisted of computing correlation coefficients between the standard deviation of projected scores and selected variables. For example, a correlation was computed between the standard deviations among projected scores for each child (see Table 3) and variables of interest which were number of test items projected (i.e., length of test), pretest performance level, and related subtest performance. A variation of this procedure was used for handicap and extent of information reported in the child's case profile.

Data presented in Table 4 show that chronological age was generally not significantly related to variability in projected scores, but there was a significant ($p < .05$) correlation with the Bayley mental scale and the KIDS gross motor scale. On these scales, the negative correlation coefficient indicated that experts tended to be less variable in their score projections with older children. While some of the remaining observed correlations were substantially different from zero, the sample size (N=17) was too small for statistical significance. Of course, larger samples may have yielded smaller coefficients.

Another interesting finding in Table 4 deals with the number of items projected. It was thought that experts might become more variable in their projections as they projected more and more test items, the data did not support this hypothesis in the Bayley scales, but there was some support for the hypothesis in the KIDS Inventory scales. There was a significant correlation for the KIDS self-care scale ($r = .58$), and the correlation ($r = -.47$) for the KIDS fine motor scale was almost significant. The pattern of observed

Table 4

Correlation Coefficients
for Selected Ranked Variables

Correlated Variable	Expert Variability in Projected Scores[b]					
	Bayley		KIDS Inventory			
	Mental	Motor	Gross Motor	Fine Motor	Cognitive Language	Self-Care
Chronological age	-.52*	-.34	-.51*	.29	-.14	.20
Number of items projected[a]	-.37	-.10	-.25	-.47	-.16	.58*
Pretest performance (raw score)	.54*	-.40	(no pretest data available)			

*Significant at the .05 level (p < .05, r = .48).

[a] Defined as the total number of items projected from the basal or entry level for a given child to the average number of the ceiling level items projected.

[b] Number of items within each subtest are as follows: Bayley mental - 163, Bayley motor - 81, KIDS gross motor - 77, KIDS fine motor - 70, KIDS cognitive/language - 112, KIDS self-care .64.

correlations between expert variability and number of items projected may reflect the psychometric status of the KIDS Inventory as compared to the Bayley. Hence the number of items projected tended to be associated with expert variability in the KIDS Inventory, but there was no such association evident in the more sophisticated Bayley scales.

The data in Table 4 also indicate there was a significant correlation between expert variability and pretest performance level on the Bayley mental scale (r = .54). The postive coefficient indicated that experts became more variable as children attained higher mental development. However, the trend, though nonsignificant, was reversed in the Bayley motor scale ($r = -.40$). Unfortunately, pretest KIDS Inventory scores were unavailable for similar study.

Another question of interest was whether or not there were significant relationships among expert variability in projected scores and subscales within the Bayley and KIDS Inventory. In other words, if there tends to be high expert variability for given children in the Bayley mental scale, does there also tend to be high expert variability for given children in the Bayley motor scale? Correlational analysis showed there were no such relationships in projected scores. The correlation between Bayley mental and motor scales in terms of expert variability was .23. The correlations among KIDS Inventory scales (Table 5) also revealed nonsignificant correlation coefficients, which ranged from -.12 to .35.

Table 5

Correlations Among KIDS Inventory Scales in Terms of Variability in Projected Scores

	Gross Motor	Fine Motor	Cognitive/ Language	Self-Care
Gross Motor	1.00	.35	.13	-.12
Fine Motor		1.00	.26	.02
Cognitive/language			1.00	.10
Self-Care				1.00

[a] Similar correlations computed between the Bayley mental and motor scores was .23. (For $p < .05$, $r = .48$)

The final variables of interest relative to expert variability were handicap and extent of information reported in the child's case profile. Since these variables were a little difficult to quantify, the analysis adopted was a variation of correlational analysis. The basic procedure was to select those children for whom experts were consistently variable and less variable in their projected scores. Standard deviations of projected scores (as reported in Table 3) were plotted in a frequency distribution, and then one or more children were selected from each end of the distribution who were termed outliers in each scale. Outliers were determined by visual inspection, and the outcome of this procedure are reported in Table 6.

Table 6

Results of Outlier Analysis

Scale	Expert Variability	
	Low	High
Bayley Mental	1, 2[a], 3[a], 5, 8[a]	4[b], 11[b]
Bayley Motor	6, 7, 13	5, 10[b], 11[b], 15
KIDS gross motor	2[a], 6, 7, 8[a], 12, 16	1, 10[b], 11[b]
KIDS fine motor	8[a], 9, 17	4[b], 10[b], 11[b], 14[b], 16
KIDS cognitive/lang.	3[a], 5, 13	2, 4[b], 11[b], 14[b]
KIDS self-care	1, 2[a], 3[a], 9, 15	6, 14[b]

[a] Children number 2, 3, and 8 were selected as having consistently high expert agreement (i.e., low expert variability).

[b] Children number 4, 10, 11, and 14 were selected as having consistently low expert agreement (i.e., high expert variability).

Table 6 shows that children numbered 2, 3, and 8 were in the outlier group for low expert variability for at least three of the seven scales, and children numbered 4, 10, 11, and 14 were in the high expert variability group for at least three of the seven scales. In other words, experts consistently agreed in their projections for children numbered 2, 3, and 8, and they consistently disagreed in their projections for children numbered 4, 10, 11, and 14. Thus, the above procedure, yielded two small groups of children for further study in terms of handicap and extent of information reported in the case profile.

Review of the case profiles of the above two groups of children revealed no differences of discernible consequence between the groups. Apparently, expert variability was not related to handicap involvement nor extent of descriptive information, but this conclusion may be misleading since there was relatively little variation among sampled case profiles in terms of information available, and almost all handicapping conditions were in the severely handicapped range.

The data in response to question 1.4 about variables associated with expert variability in projecting scores suggested that there may be some types of children for whom experts are likely to agree more in projected scores. Results from the current sample of 17 children and 4 experts indicate that chronological age and developmental level are the variables most likely to influence expert variability. Psychometric sophistication of the instrument used for projecting scores is also throught to have some association with variability of projected scores. While it seems reasonable that some children would present a more difficult score projection task, the results of the study provided only a first level investigation into this issue.

2.1 Was there significant improvement in actual pre-post scores of sampled children?

While there were projected posttest scores for 17 Project KIDS children, there were actual pre-post Bayley mental and motor scores for only 12 and 11 children, due to attrition in the sample over the 12-month interval. As noted previously, no actual pre-post KIDS Inventory scores were available.

A repeated-measures ANOVA was used to test the significance of pre-post improvement on the Bayley mental and motor scales. Table 7 gives the results of the ANOVAs, and Table 8 presents means and standard deviations for actual test scores.

Table 7

Repeated-Measures ANOVAs for
Actual Pre-Post Bayley Raw Scores

Scale	Source	df	MS	F	p
Mental	Pre-post	1	19795.04	66.21	< .001
	Residual	11	163.04		
Motor	Pre-post	1	2978.91	41.08	< .001
	Residual	10	72.51		

Inspection of Tables 7 and 8 clearly shows that sampled children made highly significant pre-post improvement. Conversion of average pre-post raw scores to developmental ages indicated an average gain of about eight and one-half months on the mental scale and about seven months on the motor scale. The above results do not necessarily indicate that the observed improvement was due to project intervention, but they do show that sampled children gained while in the project.

Table 8

Means and Standard Deviations
for Actual Pre-Post Bayley Raw Scores

	Mental		Motor	
	pre	post	pre	post
Mean	70.58	113.00	24.18	47.45
Standard deviation	36.27	42.46	16.59	21.21

2.2 What were the characteristics of theoretical control scores?

Questions 1.1 through 1.4 of this report spoke to psychometric vigor of scores projected by the panel of experts, and these data supported the feasibility of adopting a projected score as a control score for comparison to actual posttest scores. Since pre-post KIDS Inventory scores were unavailable, the theoretical control strategy necessarily relied on projected Bayley scores.

Reliability analyses (see question 1.1) showed that the reliability of the average Bayley score projected by four experts was a .81 for the mental scale and .95 for the motor scale. This appeared to justify adoption of the average projected score for each child as the theoretical control score. On the mental scale, control scores ranged from 43.7 to 140.0, and the mean and standard deviation were 104.1 and 33.1. On the motor scale, control scores ranged from 31.5 to 141.3, and the mean and standard deviation were 40.7 and 14.6.

2.3 How did theoretical control scores compare to actual posttest scores?

A repeated-measures ANOVA was adopted as the statistical model of choice to test the hypothesis of no significant difference between actual posttest

Bayley performance and theoretical control performance on the Bayley. This model might at first appear inappropriate, since the design of the theoretical control strategy can easily seem to resemble the traditional independent groups model consisting of two groups, one control and one experimental. In such a design, one would employ the one-way ANOVA for independent groups. However, the repeated-measures model is clearly the appropriate statistical model, since the actual posttest score and the theoretical control score are dependent upon the same observational unit (i.e., the child).

The results of the repeated-measures ANOVA yielded nonsignificant differences ($p < .05$) between the control and actual posttest Bayley scores. However, the observed difference for the motor scale approached significance at the .05 level. Table 9 gives the ANOVA summaries and Table 10 presents relevant means and standard deviations.

The data in Tables 9 and 10 show that the trend in observed differences supports a hypothesis favoring project intervention in that the actual scores were higher than the control scores. While the observed differences were not significant at the .05 level, motor scale improvement was significant at the .10 level. Hence, one would reject the null hypothesis of no treatment effect for motor scores at the 90 percent confidence level, but the 95 percent level is traditionally the lowest acceptable confidence level.

The data in Table 10 show that the observed difference between control and actual average scores on the motor scale was 6.19 units. Even though this difference may appear to be rather large, one typically interprets the magnitude of differences relative to group variability. The best way to do this is to form a ratio comparing the observed difference to the sample standard deviation. In the case of motor scores, this ratio equals .37.

Table 9

Repeated-Measures ANOVA Comparing Theoretical Control and Actual Posttest Bayley Raw Scores

Scale	Source	df[b]	MS	F[a]	p
Mental	Control-actual	1	429.26	2.78	NS
	Residual	11	154.27		
Motor	Control-actual	1	210.80	3.87	<.10
	Residual	10	54.44		

[a] df = 1, 11, F = 4.84, p < .05; df = 1, 10, F = 4.96, p < .05; df = 1, 11, F = 3.23, p = < .10; df = 1, 10, F = 3.28, p < .10.

[b] There were actual posttest scores available for 12 children on the mental scale and for 11 on the motor scale.

Table 10

Means and Standard Deviations for Theoretical Control and Actual Posttest Bayley Raw Scores

	Mental		Motor	
	Theoretical Control	Actual	Theoretical Control	Actual
Mean	104.54	113.00	41.26	47.45
Standard Deviaiton	33.20	42.46	15.19	21.21

(The sample standard deviation was obtained by averaging variances for the control and actual scores.)

The first reaction to the above ratio of .37 may be that the observed difference was not even of practical significance, let alone statistical significance, since a difference of one-third a standard deviation or less is usually considered trivial. However, consideration should be given to the special nature of the subject population in the study. In the usual academic setting, one-third standard deviation's difference would mean X number of additional test items correct. These items would likely pertain to such things as the correct answer to 'how much does 4 x 12 equal' or 'what is the largest continent' or 'what is the correct spelling of cat?' In the case of Project KIDS, Bayley test items dealt with such things as responding to a verbal request or standing unassisted. Thus, one can see that the usual method of ascertaining practical significance may not apply when test items call for behaviors of greater practical significance.

As noted earlier the average Bayley control score on the motor scale was 41, and the average actual score was 48 (rounded upward). The behaviors contained in the additional seven motor items are: walking with help, sitting, playing pat-a-cake, standing alone, walking alone, standing up, and throwing a ball.

The implication of the foregoing discussion is that getting an item such as 'standing up' correct has a great deal more practical significance than an item such as 'how much is 4 x 12?' Hence, the practical significance of the observed difference in motor scores between the theoretical control and actual scores is probably much greater than the computed ratio of .37 indicates. (Of course, the additional behaviors mastered by children would vary considerably. The above behaviors contained in the consecutive items between control and actual score averages were presented only as examples). Porter, Schmidt, Floden, and Freeman (1978) have also recently called for the interpretation of effect size in terms of substantive program goals and substantive characteristics of the criterion test.

The above line of thought leads one to consider a 90 percent confidence level for rejection of the null hypothesis of no treatment effect in the motor domain. This researcher recommends adoption of a 90 percent confidence level (or .10 significance level) in this particular situation because of the extremely small sample size (N=11, motor) and because of the practical significance of differences between means which are small in a statistical sense. As is well known, the F-test can be very powerless in the case of small sample size and small differences between means and thereby unlikely to detect differences which would be statistically significant with a larger sample or difference. One simple way to increase statistical power is to reduce the confidence level, and such a procedure appears warranted in this situation.

The analyses in response to question 2.3 indicate a significant ($p < .10$) effect favoring Project KIDS intervention in terms of improved progress in motor development. In other words, the actual posttest motor scores of project children were significantly greater than predicted by the panel of experts. Hence, one can conclude that observed progress in the motor domain was greater than would have been expected from normal growth and maturation.

While similar logic might be applied to the observed gains in mental scores, it was inappropriate to do so, since differences between mental actual and control means were not statistically significant ($p < .10$). These results might appear confusing in that one might expect intervention to be more effective in the mental domain than in the motor domain. At this time there is no clear explanation available except that expert panel members may have had a similar expectation and thereby projected more conservatively in the motor domain. Of course the basic assumption of the study was that exprectations (i.e., projections) of panel members were valid.

One interesting issue raised by the foregoing discussion is whether or not the same rationale should be applied to the test for an expert main effect (see question 1.3, page 14) and the answer would seem to be yes. Inspection of Table 2 (page 16) shows that the ratios of the range of mean for experts to group variability were generally quite small for all subtest except the KIDS cognitive/language and the KIDS self-care scales (34 percent and 42 percent respectively). The F-ratio does not approach significance for the cognitive/language scale, but it is significant at the .10 level for the self-care scale. Visual inspection of expert means shows that experts three and four were comparable but that expert one differed substantially from expert three.

Conclusions

The following lists the conclusions and results of the study:

1. Internal consistency reliabilities of projected scores for both the Bayley and KIDS Inventory ranged from .81 to .95, and these reliabilities were considered to be very good, especially in light of the developmental nature of the theoretical control strategy. The following gives the reliability coefficients for each test scale:

Bayley mental	.81
Bayley motor	.95
KIDS cognitive/language	.91
KIDS gross motor	.90
KIDS fine motor	.93
KIDS self-care	.87

2. The study did not empirically investigate the validity of scores projected by the expert panel. However, the credentials of experts on the panel were judged to have been sufficient to warrant a fair degree of confidence in projected score validity. Additionally, the reported content validity of both the Bayley and KIDS Inventory were sufficiently high, and this factor would logically lead to increased validity in projected scores.

3. There were no significant differences among experts in projected scores except for the Bayley self-care scale ($p < .10$). On all other scales, any given expert was not significantly higher or lower than another in his or her score projections. (For the rationale behind adoption of the 10 percent significance level, see conclusion number six.)

4. As expected, there was variability in scores projected by experts for any given child, and the extent of projected score variability differed across children. In other words, experts varied more in projecting scores for child A than for child B, for example. While it seems reasonable that some children would present a more difficult score projection task, the results provided no definitive answer as to which factors might influence expert variability. Chronological age and developmental level were seen to have some association with expert variability, but handicap, number of test items projected, and extent of available information were relatively independent of expert variability.

5. Sampled children made highly significant improvement in developmental progress during the 12-month pre-post observation period. On the

average, children gained eight and one-half months on the Bayley mental scale and seven months on the Bayley motor scale.

6. Comparison of theoretical control scores to actual Bayley scores showed that the Project KIDS children performed significantly ($p < .10$) better than the control in the motor domain, thereby indicating that motor gains were greater than would have been expected without project intervention. The reader should not discount this significant finding, even though the significance level is not at the usual .05 or .01 level. Rather, project intervention should be considered to have made a significant difference just as if the significance level had been .05 or .01. (The rationale for the above conclusion is based on two statistical concepts known as 'power' and 'practical significance'. Power has to do with the probability of finding a significant difference when in fact project intervention is truly effective. If the power of a statistical test is low, there will not be much chance of finding a significant difference between groups even when the treatment is effective. Practical significance applies to the practical meaning of any observed significant difference. In some cases, power can be so great that a very small difference of no practical significance can be statistically significant. An important point is that small differences require considerable power to be statistically significant. In the case of the Project KIDS evaluation, the difference between the control and treatment scores was relatively small, about one-third a standard deviation, which would usually be considered to have no practical significance. However, Project KIDS served a unique population, unlike that encountered in the usual public school evaluations of reading and math programs. While test items measuring impact of a language arts program would test skills in word recognition, spelling, and so

forth, test items measuring impact of Project KIDS tested skills in such basic behaviors as sitting and standing. Thus, improvement in even a few test items would seem to have much more practical significance than in the case of the usual language arts program. Thus, the statistical test in the case of Project KIDS needed to be relatively powerful in order to detect even a small difference between the control and treatment groups. The 90 percent confidence level was adopted to increase power and to detect any observed difference which would have been significant given a larger sample size in the project evaluation.)

References

Bayley, Nancy. Manual: Bayley Scales of Infant Development. New York: The Psychological Corporation, 1969.

Carter, J. L. Project KIDS student progress report, Research Report No. SP78-105-57-08, Department of Research, Evaluation, and Information Systems, Dallas Independent School District, Dallas, Texas, 1978.

Carter, J. L. The assessment of competency levels of parents. Presentation given to the annual meeting of the Council for Exceptional Children, Kansas City, 1978.

Curtis, W. Scott and Donlon, Edward T. The development and evaluation of a video-tape protocol for the examination of multihandicapped deaf-blind children, Final Report to the Bureau of Research, Division of Handicapped Children, United States Office of Education, Project Number OEG-0-9-442134-2764(032), 1972.

Goulet, L. R. Longitudinal and time-lag designs in educational research: An alternate sampling model. Review of Educational Research, 1975, 45, 505-524.

Project KIDS. KIDS Inventory of Development Scale. Dallas: Dallas Independent School District.

Macy, Daniel J. and Carter, Jamie L. Plan A and the non-LD student, Research Report No. 75-595, Department of Research and Evaluation, Dallas Independent School District, Dallas, Texas, 1974.

Macy, D. J. Determination of parent competencies through needs assessment procedures. Presentation given to the annual meeting of the Council for Exceptional Children, Kansas City, 1978.

Porter, A. C., Schmidt, W. H., Floden, R. E., and Freeman, D. J. Practical significance in program evaluation. American Educational Research Journal, 1978, 15, 529-39.

Simeonsson, Rune J. and Wiegerink, Ronald. Accountability: A dilemma in infant intervention. Exceptional Children, 1975, 41, 474-481.

Stricklin, James L. The Psycho-social Index: A Systematic Method of Case Data (2nd ed., revised). Cape Town, South Africa: Unviersity of Cape Town Press, 1974.

Turner, R. M. Individualized competency based programming for parents. Presentation given to the annual meeting of the Council for Exceptional Children, Kansas City, 1978.

Winer, B. J. Statistical Principles in Experimental Design (2nd ed.). New York: McGraw-Hill Book Company, 1971.

Appendix

PROJECT KIDS

THEORETICAL CONTROL GROUP STUDY

DIRECTIONS TO EXPERTS

The following details the content of your materials packet and describes procedures for making and recording your projections.

A. PROJECT KIDS EVALUATION DESIGN

This document is included only for your information about the project evaluation in general and more specifically about the design of the theoretical control group study (see pp. 17-25 for discussion of the theoretical control strategy).

B. INDIVIDUAL CASE REPORTS (N=17)

There is one case report for each of 17 children. Each case report contains the following information:

1. Subject's Summary Sheet (one page)
2. Project KIDS Referral Form (one page)
3. Family Information Form and Home Assessment (usually 10 pages)
4. Denver Developmental Screening Test (one page)
5. Bayley Scales of Infant Development (variable pages) (These give the Bayley pretest results.)
6. Medical Data (variable pages)
7. University Affiliated Facility Summary (variable pages)

C. BAYLEY RECORD FORMS

This includes 17 copies of the Mental Scale Record Form and the Motor Scale Record Form for the Bayley Scales of Infant Development.

Directions for recording your projections on the Bayley Record (both Mental and Motor) forms:

1. Note the PROJECTED CA on the Subject's Summary Sheet. This is the age the child would be at twelve months past the time of pretesting with the Bayley (pretest results for each child are included in the individual case report - item B.5.)

2. Based on the total relevant information contained in each child's individual case report, project what you think each child's performance will be on the Bayley Mental and Motor Scales at time of the PROJECTED CA. Assume that there has been no special (remedial, therapeutic, educational, clinical, etc.) interventions into the child's life. In other words, how might the child have progressed had he lived in his family during the twelve month period without benefit of services other than those his or her family could provide? Please assume that any medical intervention necessary for life support which was present at the start of the twelve month period was continued or that any life support intervention which might become necessary during the twelve months was made available to the child.

3. Record your projections (i.e., check the items "P" or "F") on the Bayley Record Forms for each child as though you had actually administered the Bayley to each child. Use the pretest ceiling level as the basal level for the starting point in "administering" your projected Bayley.

   Example: if a child's ceiling level on the pretest Mental Scale is item 47, then the basal level for beginning the projected Mental Scale would be item 47. Please note that the basal item for the projected Bayley would be the last item passed prior to termination of the pretest Bayley.

   Please follow the Bayley Manual specifications for determining the ceiling level of your projected Bayley (i.e., a criterion of 10 successive items failed on the Mental Scale and 6 successive items failed on the Motor Scale; Manual for the Bayley Scales of Infant Development, the Psychological Corporation, 1969, p.29).

   It is not necessary for you to compute a composite score (either raw or normed conversion) as we will do that here in our office.

D. KIDS INVENTORY OF DEVELOPMENT

This includes one copy of the KIDS Inventory of Development manual and a copy of the KIDS Inventory of Development record form for each child (N=17).

1. KIDS Inventory of Development - MANUAL

   This document describes the development and rationale behind the Inventory, as well as giving a description and directions for administering individual items in the Inventory.

2. KIDS Inventory of Development - RECORD FORM (N=17)

   The RECORD FORM contains four major developmental areas: gross motor (GM), fine motor (FM), cognitive/language (CL), and self-care (SC). Within each area, Inventory items are grouped according to chronological age intervals.

   Directions for recording your projections on the KIDS Inventory of Development:

   a. Use the same PROJECTED CA (see Subject's Summary Sheet) that you used for each child in making the Bayley projections:

   b. Project what you think each child's performance on the Inventory would be at the time of the PROJECTED CA (which is 12 months past the age of Bayley pretesting). Use the same rationale and assumptions used in making the Bayley projections (see item C.2.).

   c. Record your projections for each child on an Inventory RECORD FORM as though you had actually administered the Inventory to each child. (Note that your materials packet does not contain any pretest Inventory results.)

      When "administering" the Inventory, testing should be initiated at the level at which the child is expected to achieve. A basal is established when a report of P (pass) is recorded for all items within one age interval before the first failure. Testing should continue until the child has failed all items within one age interval, indicating a ceiling.

      Observe that projections must be made within each of the four developmental areas (GM, FM, CL, and SC) of the Inventory. There is no need for you to figure a composite Inventory score for the child.

* * * * RETURN OF MATERIALS * * * *

TIME: At your earliest convenience.

PLACE: Return materials to

Dr. Ruth Turner
Assistant Director-Special Education
Dallas ISD
3700 Ross Ave.
Dallas, Texas 75204

WHAT: It is only necessary to return the record forms for each child.

Be sure that you have placed the child's first name on each record form for each child. Please group the record forms by child according to this order within each child: Bayley Mental, Bayley Motor, KIDS Inventory. (Please include some identifying information in your return correspondence, so that we know for sure which record forms are from which expert.)

The individual case reports may be retained if desired (as well as the KIDS Evaluation Plan and KIDS Inventory). You may wish to refer back to the case reports when the results of the study become available.

If you have any questions, please feel free to call Dr. Daniel Macy, Department of Research & Evaluation, Dallas ISD (phone 214 - 321-2667).

ANOVA SUMMARIES

Test	Source	df	MS	F	p
Bayley Mental	Between subjects	16	2804.50	6.25	< .001
	Within subjects	51	448.79		
	Between experts	3	67.39	0.14	NS
	Residual	48	472.63		
Bayley Motor	Between subjects	16	750.51	22.99	< .001
	Within subjects	51	32.65		
	Between experts	3	32.57	1.00	NS
	Residual	48	32.65		
KIDS cognitive/ language	Between subjects	16	403.53	12.22	< .001
	Within subjects	51	33.01		
	Between experts	3	46.76	1.45	NS
	Residual	48	32.16		
KIDS gross motor	Between subjects	16	581.36	10.21	< .001
	Within subjects	51	56.93		
	Between experts	3	89.08	1.62	NS
	Residual	48	54.92		
KIDS fine motor	Between subjects	16	445.75	14.83	< .001
	Within subjects	51	30.05		
	Between experts	3	12.31	0.40	NS
	Residual	48	31.16		
KIDS self-care	Between subjects	16	180.52	8.42	< .001
	Within subjects	34	21.45		
	Between experts	2	56.61	2.94	< .10
	Residual	32	19.25		

Reliability Computations for Expert Panel
(See question 1.1)

The following presents the reliability computations for determining the intraclass correlation coefficient derived from the analysis of variance model. The basic procedure was to compute an unbiased estimate of theta (Ø), the term used to denote the ratio of true score variance to error score variance, where

$$Ø = \frac{MS_b - (X)\ (MS_w)}{k\ (X)\ (MS_w)} \quad \text{where}$$

k is the number of experts,
MS_b is the mean square between subjects from the ANOVA, and
MS_w is the mean square within subjects from the ANOVA, and

$$X = \frac{n\ (k-1)}{n\ (k-1)\ -2} \quad \text{where n is the}$$

number of subjects.

The above term, Ø, was then used to compute the intraclass correlation coefficient as follows:

$$r_1 = \frac{Ø}{1 + Ø} .$$

The above equation gives the reliability estimate for a single expert. The Spearman-Brown prediction formula was used to obtain the reliability estimate for the average of all four experts,

$$r_4 = \frac{4\ Ø}{1 + 4\ Ø} .$$

For a more detailed description of the above procedure, see Winer (1971, pp. 283ff.). The ANOVA model also permits adjusting the error variance to partition out that variance attributable to "expert" main effects. This was not done since there were no significant expert main effects in any of the test scores (refer to ANOVA summaries on previous page).