DOCUMENT RESUME

ED 077 968 TM 002 777

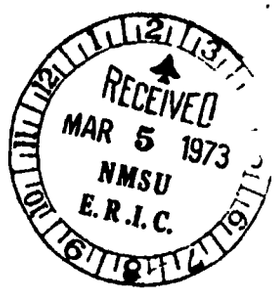
AUTHOR	Rost, Paul; And Others
TITLE	A Model for Evaluating Title 1 Programs.
INSTITUTION	Albuquerque Public Schools, N. Mex.
PUB DATE	Nov 72
NOTE	31p.; Paper presented at the meeting of the Rocky Mountain Educational Research Association (Las Cruces, New Mexico, November 16-17, 1972)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Elementary Grades; Evaluation Criteria; *Evaluation Methods; Evaluation Needs; Federal Programs; Kindergarten; *Models; *Program Evaluation; Research Design; *Research Problems; Speeches
IDENTIFIERS	*Elementary Secondary Education Act Title I; ESEA Title I

ABSTRACT

Albuquerque's Title I evaluation staff is in the process of generating a comprehensive local evaluation design because it considers the federally required product evaluation unsatisfactory. The required mean-gain comparisons were extended beyond the dimension of program to the dimensions of school, grade, and Title I instructor. This evaluation effort is an attempt to sort out elementary program variables and at the same time confront the problem of inadequate controls. A summary of the first efforts at process evaluation is also made, and evaluation results for 1971-72 are given. Future evaluation will involve process evaluation for program activities and product evaluation to search for effective variables. Problems in implementing the evaluation design include: (1) lack of understanding and acceptance by school and program personnel of the role of evaluator; (2) the difficulty of persuading program personnel to specify measurable objectives and to permit adequate controls; (3) the fact that some of the most relevant variables, such as teacher effectiveness, are taboo; (4) inadequate programming and programmer time for data processing; and (5) inadequate dissemination of findings. (KM)

ED 077968

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY



A Model for Evaluating Title I Programs

Paul Rost
Jerry Seidenwurm
Andrea Vierra

Title I Evaluation Staff, Albuquerque Public Schools

TM 002 777

Paper delivered at the Rocky Mountain Educational Research Association, Las Cruces, New Mexico (New Mexico State University), 16-17 November 1972.

FILMED FROM BEST AVAILABLE COPY

A Model for Evaluating Title I Programs

Introduction. The purpose of this paper is to describe the design used in evaluating Title I projects in the Albuquerque Public School system; to present the major results of the most recent evaluation; to discuss the problems that arise in attempting to implement an evaluation design in a public system; and to present the design revisions that we think may solve some of those problems.

Before we can discuss Title I program evaluation, however, a brief description of the programs themselves is necessary.

Title I Programming in Albuquerque. Title I law authorizes the appropriation of federal funds for improving the academic skills of educationally disadvantaged children living in low-income areas. For Title I purposes, low-income means roughly $3000 or less annually for a family, of whatever size. In Albuquerque, a Title I low-income area is a public elementary school attendance area where a sixth or more of the families live on less than $3000 a year. Twenty-nine public elementary school attendance areas qualify; twenty-three of these actually participate in Title I programs. In most of the participating areas, at least a third of the families are low-income. Four parochial schools are also involved with Title I, since some of their students live in these low-income areas.

Title I law requires that a comprehensive needs assessment be conducted in those attendance areas receiving Title I funds so that programs can be tailored to meet felt needs. Needs assessments involving students, parents, teachers, and administrators have been conducted over the last two years in Albuquerque's Title I attendance areas. The results indicate two primary needs: kindergarten (Albuquerque does not have public kindergartens) and development of basic academic skills. The two basic Title I programs in Albuquerque are, in fact, kindergarten for five-year-olds and reading for first- through sixth-graders. The possibility of adding a math program for the same grades is being considered.

After participating schools are selected and specific needs are assessed, eligible children within the schools must be identified. The number of children who receive Title I services is conditioned by the amount of funding. At least half again as much as what the State and local school systems spend to educate each child must be added by Title I for each participating child. Since it costs about $600 a year to educate a child in the Albuquerque Public School system, Title I in Albuquerque must add at least $300 a year per participant. Albuquerque's level of Title I funding in 1972-73 permits service to about 3950 children.

Title I law mandates identification of educationally disadvantaged children in eligible schools through testing with a nationally standardized instrument. In Albuquerque, the Stanford Achievement series is used. Exact cut-off scores vary from year to year, since the number of children who receive services is conditioned by the amount of funding. In 1972-73, for example, sixth graders in eligible Albuquerque schools had to have scored below 3.2 grade equivalents on the Stanford reading sub-tests at the end of the fifth-grade in order to qualify for Title I services.

Title I programming in Albuquerque is either instructional or supportive. The instructional programs are of two basic types: kindergarten and elementary language. All twenty-three public Title I elementary schools have at least one full-time Title I language specialist and at least one Title I kindergarten teacher-and-aide team. The four parochial schools, which have a relatively low number of eligible children, share two reading teachers.

It has been the policy in Albuquerque to admit all age-eligible children in Title I attendance areas to kindergarten. The per cent of kindergarteners who are educationally disadvantaged is then determined by administering the Stanford Early School Achievement Test (SESAT), Level I, in early October. The results of the October '72 testing indicate that eighty- nine per cent of the Title I kindergarten

children in Albuquerque are educationally disadvantaged, using a criterion score of less than the fiftieth percentile on the Environment or Aural Comprehension subtest. The children who do not qualify as educationally disadvantaged according to this criterion are permitted to remain in Title I kindergartens for two reasons: first, space is available, and second, they serve as appropriate models for the disadvantaged majority. Of the 3950 children who are receiving Title I services in Albuquerque in 1972-73, about 1300 are in kindergarten programs. The remainder are eligible for elementary language programs.

In order to designate children eligible for the next year's elementary language programs, the Albuquerque Title I office administers the Stanford Achievement Test (SAT) reading subtests to all children in Title I schools in the spring. The number of children for whom budget permits service is then portioned out among the grades, and the lowest scoring children at each grade level are selected. Master Lists of eligible children at each grade level in each school are prepared and distributed to the schools. Title I personnel in the schools must then restrict their efforts to Master List children. Children who were not tested or whose test scores are suspect may be added to the Master List if subsequent testing by reading specialists establishes that their need is as great as that of other Master List children in their grade.

Each of the basic instructional programs has an overall objective. For the kindergarten program, it is to prepare Title I children to function more successfully when they enter the regular school program, as measured by the Stanford Early School Achievement Test. The overall objective of the elementary language programs is to slow the rate at which Title I children fall behind in reading achievement, as measured by the reading subtests of the Stanford Achievement Tests. Since engineering an achievement increase is traditionally difficult with these children, an objective

of five months mean gain over a seven-month instructional period seemed realistic; although this objective does not catch the children up, it does represent a substantial decrease in the rate at which they traditionally fall behind.

In addition to the two instructional programs, three support programs are funded in Albuquerque by Title I: counseling, health and nursing services, and tutoring. The purpose of the support programs is to contribute to the objectives of the instructional programs. Only children who are eligible for a Title I instructional program are eligible for support services; but not all eligible children receive a substantial amount of support service. For the most part, support personnel screen eligible children and render service where they find a need and as they have time available.

Albuquerque's Title I Evaluation Design & Last Year's Results: Product Evaluation. The Title I evaluation design in Albuquerque is conditioned, first, by State and federal requirements. Although Washington allows five per cent of local Title I budgets to be spent on evaluation, only the grossest kind of product evaluation of program impact is required. The Albuquerque Title I office must simply submit, at the end of each program year, the mean pre- and post-test scores and score ranges, on a standardized instrument, for two groups of children in each separately funded program: 1.) children who were eligible for Title I services and 2.) children who attended Title I program schools but whose pre-test scores were too high for service eligibility.

The remainder of the local Title I evaluation unit's responsibility involves rules and regulations. Tabs must be kept on whether Title I law and guidelines are being adhered to. It is generally understood that loss of funds by local programs is less related to failure to achieve instructional objectives than to two other causes: if federal auditors find individual local agencies in violation of Title I law and guidelines, they may be required to pay back misspent funds; and

if Congress cuts appropriations, all local agencies suffer equally. In general, however, any local agency that follows the rules is assured of getting its share of the Title I funds, whatever its instructional program results.

In terms, then, of the mean results by program required by the State and federal agencies, evaluation of Title I programs in Albuquerque for the past year (1971-72) was probably more favorable than unfavorable. Kindergartens registered a mean gain for the year of four percentile points on the SESAT--up from a pre-test percentile rank (in terms of national norms) of 16 to a post-test rank of 20. No local control group of the type specified by the State office was available, all public kindergarten children in Albuquerque receive Title I services.

Of the seven separately funded language programs, three obtained or exceeded the objective of five months mean gain on the SAT reading subtests; two others came very close to obtaining the objective (see Table 1, page 7). However, when the treatment children were compared with the control children, the controls showed greater gains in four out of seven cases, two of them substantial.

The use of the term "control" in this context is, however, questionable. Comparison children, by State mandate, are children in Title I schools whose test scores are too high to qualify them for Title I service. It seems reasonable to expect greater gains from these children. For that reason, we also compared treatment children with children in Title I schools who were eligible for service on the basis of their low test scores but who, for a variety of reasons, did not receive service. In these comparisons, the controls showed as great or greater gains in six out of seven cases, two of them substantial.

We are not, at this point, at all sure that the eligible non-served group of controls represents much of an improvement over the State required non-eligible group. Although the non-served controls, like treatment children, were eligible for

Title I service, entry means for the controls vary by as much as five months from treatment entry means, sometimes higher, sometimes lower.

This seems to be more a function of differential service to children in the various grades, however, than of any selection for more or less needy children. In reading program E, for example, the entry mean for served children was 1.51, with a mean gain of .37. Eligible non-served children, on the other hand, had an entry mean of 2.11 and a mean gain of .51. It is tempting to conclude that program E is helping the neediest children and that these children's slower progress can be explained by their greater need.

The fact, however, is that personnel in reading program E elected to work only with first through fourth graders. The treatment group, then, does not contain any fifth graders, while the control group contains a disproportionate number of these children; the controls' entry mean must, obviously, be higher. When the control group means are adjusted by eliminating fifth graders, the discrepancies between treatment and control groups are sharply reduced (see Table 2). The entry mean for treatment children is now slightly higher, rather than substantially lower, and the mean gain is almost indistinguishable, rather than substantially lower.

We know that poorly matched controls are influencing the results in many of our programs. The higher entry means for treatment children in reading programs A and F reflect the fact that these programs systematically exclude first graders. In program B, the reverse happens--first graders tend to be over-represented among served children. Adjusting for these phenomena would raise control gains for A and F; for B, however, treatment gains would increase.

We do not, at this point, have the data-processing capability to more carefully match controls; and with a population of over 10,000 served and un-served children, we cannot possibly deal with this problem on a desk calculator. When hand calculations

Table 1

Pre-test means and mean gains on SAT reading subtests (for kindergarten only, substitute SESAT total battery) for 1971-72, by program. N's represent all children for whom pre- and post-test scores were available, except sixth graders. The sixth grade was excluded because the testing interval was different for these children. Kindergarten scores are reported in terms of percentile rank compared to national norms; all other scores represent grade-level equivalents.

Type of program	Served			Eligible non-served			Non-eligible		
	N	Pre $\bar{X}$	$\bar{X}$ gain	N	Pre $\bar{X}$	$\bar{X}$ gain	N	Pre $\bar{X}$	$\bar{X}$ gain
Kindergarten	576	16%	4%	Not applicable			Not applicable		
Elementary language programs									
A	210	1.86	.64	280	1.67	.51	1202	2.53	.60
B	221	1.74	.51	374	1.81	.57	1001	2.59	.58
C	23	1.97	.56	40	1.93	.57	283	2.54	.53
D	26	1.93	.35	142	1.82	.53	251	2.51	.61
E	102	1.51	.37	51	2.11	.51	107	2.29	.44
F	19	2.00	.47	39	1.86	.47	201	2.68	.46
G	129	1.68	.45	943	1.80	.54	3085	2.53	.58
Counseling	279	1.75	.45	943	1.80	.54	3085	2.53	.58
Nursing	27	1.94	.60	943	1.80	.54	3085	2.53	.58
Tutoring	137	1.81	.61	943	1.80	.54	3085	2.53	.58

Table 2

Means for reading program E, before and after more careful matching of controls.

	Before			After		
	N	Pre $\bar{X}$	$\bar{X}$ gain	N	Pre $\bar{X}$	$\bar{X}$ gain
Served	102	1.51	.37	102	1.51	.37
Eligible non-served	51	2.11	.51	22	1.46	.38

for a small sample, matching only one of many relevant variables, reveal the discrepancies presented in Table 2, we can only conclude that the overall results presented in Table 1 must be interpreted with great caution--and point out that, until we acquire the capability to match controls along a number of dimensions, our results will continue to be undependable.

Since each of the support programs is separately funded, we are also required to report mean reading-test results for the children who received support service, by program. For what it is worth, the mean gains for children receiving service exceeded the criterion of five months in two of the three support programs. However, there is so much overlap in these means that they are not, in fact, worth much. Most of the children who received counseling, for example, also received reading service; some of the counseled children also received speech; some received speech and tutoring; and so forth through all the possible combinations of service. It is impossible to say, on the basis of the mean scores required by the government, what part of the children's gain is attributable to a particular service.

Since federal requirements promote such an unsatisfactory and general kind of product evaluation, Albuquerque's Title I evaluation staff is in the process of generating a much more comprehensive local evaluation design. To the extent that staffing and data-processing constraints have permitted, some of this design has been applied to last year's data.

First, we extended the required mean-gain comparisons beyond the dimension of program to the dimensions of school, grade, and Title I instructor. The range of results for each of these three variables was much more dramatic than the range of results across programs, but the instructor variable may be the most significant of all four variables in accounting for variance. Depending on their instructor, kindergarten groups, for example, may lose as much as 18 percentile points in a year compared

to national norms, or they may gain as much as 27 points.

Grade level was another interesting source of variation. Gains are much greater in the intermediate than primary grades--.74 mean grade equivalents gain (seven months gain compared to national norms) for fourth and fifth grades, as opposed to .31 (three months gain) for first through third grades. Within the intermediate grades themselves, gains are better in the fourth grade (with a mean gain of nine months) than in the fifth (with a mean gain of five months). No intentional program variables (such as intent to concentrate services at the intermediate level) account for this variance.

Reference back to the available control groups, however, produces uniform results: even in grades where Title I gains were good, eligible children who were not served gained at least as much as those who were served. The instructor variable was the only exception to this rule. Seven reading instructors obtained a mean gain with served children that was at least a month greater than the gain of non-served eligibles in their schools. And hand calculations indicate that controlling for the grade differential between treatment and control groups would not, in these seven cases, cut into the treatment children's greater gains.

The overall objective of the kindergarten program is to increase children's success in regular school programs; therefore, in addition to measuring gain on the SESAT, which predicts success in elementary school, we are also attempting to measure success in elementary school directly. Last spring's testing in Title I schools included first graders who were in Title I kindergartens in 1970-71. A five-school sample shows a slight advantage on the SAT reading subtests for these children. Their mean grade-equivalent score was one month higher than that of first-grade children who did not attend kindergarten.

In addition, early in the 1972-73 school year we asked first-grade teachers in all Title I schools to list the ten students in their current classes who were best prepared for school in terms of the following criteria: listening skills and attention span; social skills; motor skills; language facility and readiness; and adjustment to school. We did not specify our interest in kindergarten; we simply told the teachers that we needed these baseline data for a longitudinal study. Preliminary analysis of a five-school sample again shows a slight advantage for kindergarten. The per cent of kindergarten children on the best-prepared lists (seventy-nine per-cent) is somewhat higher than the per cent of kindergarten children in the first-grade classes as a whole (seventy-three per cent). Relatively more kindergarten than non-kindergarten children, in other words, made the list.

Kindergarten post-test scores for these "best prepared" first-graders tend to support SESAT's validity as a readiness measure. Children selected by their teachers as best-prepared for first grade also scored well above the mean on tests administered at the end of kindergarten. This readiness, however, seems to be a quality that the best-prepared children bring to, rather than learn from, kindergarten: their kinder-garten pre-test scores, expressed as a percentile rank, are even higher than their post-test scores. This discovery, of course, only reinforces our conviction that background differences account heavily for children's differential success in school and that one year of kindergarten cannot equalize the effects of the previous five years of children's lives. Test scores and teacher evaluations of kindergarten vs. non-kindergarten children should continue to be interesting as we follow this and succeeding kindergarten classes through the grades.

Our more recent evaluation effort represents an attempt to sort out elementary program variables and at the same time confront the problem of inadequate controls. With children participating differentially in a variety of programs, the univariate,

mean-based analyses we have done to date cannot help us assess the differential contributions of instructional and support programs o reading gain or the interaction effects of these programs on reading gain; indeed, without more carefully matched control groups, we cannot claim on the basis of our previously reported analyses that any of our programs have made any contribution to reading gain. A multivariate, correlation-based technique (which does not ~ ' ntrols to establish relationships) seemed the most promising new and immediate avenue.

In addition to reading-test scores, we have records for each Title I child of the amount of service he received in each Title I program, in increments of five hours. With gain in reading achievement as the dependent variable and amounts of service in reading, speech, counseling, nursing, and tutoring as the independent variables, we performed a stepwise multiple correlation analysis; our cases were all the children who participated in Title I elementary programs in 1971-72 and for whom pre- and post-test scores were available (N=1016). The multiple R was .12: ninety-nine per cent of the variance in reading gain was unaccounted for by program variables. In terms of simple correlations between the dependent and independent variables, the two highest number values of r--.08 and .05 for counseling and speech respectively--were negative; and the r for hours of reading instruction, which should have a stronger relationship with the dependent variable of reading gain than hours of any of the less directly related services, was .03--slightly lower than nursing and tutoring at .04 each (see Table 3).

We feel that the multiple correlation is more trustworthy than previous analyses; certainly it has helped a great deal with the problems of overlapping program effects and program interaction. But we are still not entirely satisfied with our future ability to assess program impact. The problem is an interesting one and deserves some discussion.

Table 3

Correlation coefficients based on all students receiving Title I services in 1971-72; dependent variable is gain on SAT reading subtests, independent variables are amounts of service in five programs; N = 1016.

Independent variables	r
reading	.03
speech	-.05
counseling	-.08
nursing	.04
tutoring	.04
multiple R	.12

A correlation coefficient for time spent in Title I reading programs vs. reading improvement, based on all eligible children, seems to be a valid measure of reading-program impact. Children are eligible and are served because they have reading need, but some eligible children receive more reading service than others. If the program is effective, those who receive more service should improve more, and the correlation between service and improvement should be high.

But a correlation coefficient for time spent in other Title I programs vs. reading improvement, based on all reading-eligible children, is not necessarily a valid measure of program impact. The inclusion of cases for statistical analysis is still based on reading need, but actual receipt of service is now based on other needs. It becomes unreasonable to expect those who receive more auxiliary service to improve more; those who do not receive auxiliary service, because they do not need it, may very well gain as much in reading as those who need and therefore receive many hours of such service. Even if auxiliary services do increase the reading achievement of those children who need the services, the correlation between amount of auxiliary service and reading improvement for all children may be low.

This is not, in fact, an uncommon problem. The relationship between two variables--for example, counseling hours and reading gain--may be conditioned by a third, or confounding, variable--in this example, counseling need. A number of techniques exist for dealing with this situation, usually based on some method which holds the intervening variable constant. But we have been hard pressed to assign values for these intervening variables in order to hold them constant. Theoretically, Master List children receive auxiliary services based on their need for those services; in practice, however, assignment of service hours is much less precise. One counselor, to continue with the above example, only has time to work with thirty-five children and selects these children on a first-come, first-served basis; another decides to

work only with children who seem to have a particular kind of problem; while a third attempts to give equal time to all Master List children as a preventive measure. In short, the counselors themselves have only an intuitive notion as to which children need how much help.

If service personnel cannot or do not systematically assess children's need, we can neither treat need as a variable nor attempt to control for its influence. For the time being, we are relying on a seat-of-the-pants approach to the problem. If there is no correlation between hours of service in the various program; and if simple correlations between hours of service in each of the auxiliary programs and reading gain, based on only those children who got the service, are high and positive; then low program hours for children who didn't need auxiliary service may, in fact, be causing the multiple R to be spuriously low. The first of these two conditions is met: there is no correlation between hours of service in the various programs. Children, for example, who got lots of counseling help did not necessarily get lots of reading help. Any simple correlation between amount of counseling service and reading gain cannot, therefore, be attributed to overlap from some other service. However, the second condition is not met: simple correlations between hours of service in each of the auxiliary programs and reading gain, based only on those children who got each service, are consistently low. Low program hours for children who did not need auxiliary service are not, therefore, causing the multiple R to be spuriously low.

In terms of this particular analysis, then, the multiple R of .12 for all 1971-72 Title I elementary programs vs. reading gain must be accepted as substantially correct. In addition, this result tends to be supported by our previous mean-based analyses which indicate greater gain for eligible children who received no Title I service of any kind. Our conclusion must be that <u>overall</u> Title I elementary programming

in Albuquerque failed in 1971-72 to systematically improve the reading achievement of the children it served. However, we must subject our correlational analysis of overall programming to the same criticism we leveled at State and federally mandated reporting of overall means: it is the grossest kind of product evaluation and necessarily obscures the considerable variation that exists in reality.

Our extension of mean reporting across the variables of grade, school, and instructor revealed considerable variation in achievement. We were confident that similarly extending our correlation-based analysis would be even more productive, but we were limited to only one further computer run. (More will be said about this and other data-processing frustrations later.) We decided to request another multiple correlation, based only on children who participated in the most successful reading program.

Our previous analyses had indicated that this was the only instructional program in which served children gained more than controls. However, the entry mean for served children was higher than that for controls. We knew that this difference in entry means was probably a reflection of the fact that reading program A generally excluded first-graders. We could not, however, know whether this fact was also affecting the comparison between gains.

Correlational analysis based on this group of children afforded us the opportunity to check up on our admittedly shaky controls. If the correlation coefficients based on children in reading program A were higher than those based on all children, then interpretations based on available control-group data would tend to be validated. If the coefficients were similarly low, however, we would have to conclude that our inability to match controls definitely invalidates our previous results. The strength of the coefficients for children in reading program A would, of course, also be an indication as to just how much of these children's gain was attributable to what

Title I programming.

The results of this analysis are very interesting (see Table 4). Product-moment coefficients (r's) between amounts of service and gain are as low for these children as they were for all children--with one important exception. The simple r between amount of reading service and reading gain for children in reading program A is noticeably higher than the reading-service r for all children. Our earlier conclusion, based on available controls, that reading program A is the only Title I elementary program in Albuquerque that is positively and systematically affecting children's reading scores, tends to be substantiated by this analysis. We must point out, however, that even this program is not having very considerable impact. An r of .26, although substantially higher than .03, only accounts for seven per cent of the variance in children's reading gain.

Process Evaluation. We have dealt, up to this point, with only one kind of program evaluation: product evaluation, or assessment of program results in the light of program objectives. There is, of course, another kind of program evaluation: process evaluation, or assessment of program progress toward program objectives. The two kinds of evaluation require very different methodologies, and in some ways process evaluation is more difficult. The task in product evaluation is, basically, to identify, administer, and interpret valid measures of program objectives. The basic task in process evaluation is to relate program activities to program objectives. Process evaluation, in a sense, explains product evaluation; it explains why program objectives are or are not being met.

We have long felt that product evaluation alone is not very useful in improving programs: knowledge that programs are failing to meet their objectives does not necessarily imply knowledge of how those programs should be changed. On the other hand, process evaluation, which distinguishes program activities that contribute

Table 4

Comparison between students in reading program A and all elementary students. Dependent variable is gain on SAT reading subtests; independent variables are amounts of service in five programs.

Independent variables	All elementary students; N=1016 r	Students in reading program A; N=607 r
reading	.03	.26
speech	-.05	-.08
counseling	-.08	-.12
nursing	.04	.02
tutoring	.04	.02
multiple R	.12	.32

positively to objectives from activities that are neutral or negative, offers a basis for constructive program change.

We have made some attempt to get into process evaluation. In 1971-72 we contracted with an early childhood education consultant to go into Title I kindergartens and relate the activities that took place between pre- and post-testing to program objectives. The consultant's general assessment was very favorable, but her more specific comments tended to be descriptive only and limited to her linguistic specialty; their usefulness to program coordinators in attempting to improve the program is questionable. Another consultant has been hired to evaluate process in this year's kindergarten program; her contract spells out much more specifically the linking of activities to objectives. Her first report leaves open to question whether she is in fact relating activities to objectives. We will discuss her evaluation structure with her again in an effort to get at the kind of information we have been hoping for.

In addition, we have ourselves attempted some process monitoring of both the kindergarten and elementary programs. Sidestepping the problem of understaffing, which could at least theoretically be remedied, our process evaluation attempts have primarily suffered from lack of structure. Few of the programs have their own comprehensive sets of objectives, much less activities. All Title I elementary personnel and programs supposedly operate under the reading-improvement mandate; but each program and even each individual within each program are, in fact, fairly free to interpret this as they see fit. As process evaluators we are reduced to the condition of trying to decide whether each of the interactions we observe between Title I staff and students, considered on its own, bears any conceivable relationship to reading improvement.

Since Title I programming encompasses a wide range of specialties, from early childhood to oral language development, we do not always feel competent to judge the

relationship of activities to objectives. But the specialists themselves are not very reassuring or helpful in this context. Sometimes they can express the purpose of a particular activity, but at other times they cannot; sometimes they can relate their expressed purpose to reading improvement, either directly or indirectly, but at other times they cannot. Often the purported relationship--"children can't learn to read if they're unhappy"--is so general as to be meaningless.

This lack of program structure discourages any effort on our part to structure the evaluation process itself. The specialists themselves do not tell us what is important to look for. When we take matters into our own hands and make our own best judgement about what we will look for, we effectively undermine the usefulness of our results for program change. Whatever our observations, they can be conveniently dismissed by program staff--"you were looking at the wrong things."

The result of our process evaluation to date is at this point obvious. In a sense, the process evaluation has done what it was intended to do: it has explained the results of the product evaluation. The multiple R of .12, indicating that Title I programming accounts for only one per cent of the variance in children's reading gain, is only slightly higher than the R that would be expected if random variables were operative. The process evaluation indicates that randomness is not, in fact, uncharacteristic of Title I programming.

First, the various programs seem almost random in their methods of pursuing the reading objective: there is little observable consistency either from program to program within Title I overall or from individual to individual within programs. Second, the application of programs to eligible students is also effectively randomized: there are no consistently applied criteria for selecting eligible children into and out of individual programs or for determining how much time will be spent on children in programs. It is hardly surprising, therefore, that the programs do not systematically

affect the children's reading gain.

Our process evaluation also indicates the relationship between individual teacher variables (as opposed to program variables) and student gain that was suggested by our product evaluation. In each program there are individuals whose interaction with children consistently seems to promote program objectives (as well as other individuals who seem to be inconsistently effective, those who seem consistently ineffective, and even a few who probably cause children to lose rather than gain.) No other variable that we have observed (e.g., type of program, type of material, administration of program) seems to be as influential as the teacher variable. To restate a cliche, good teachers get reading gains with bad materials; bad teachers do not get reading gains even with good materials.

It would be unfair to program personnel if we failed to mention, at this point, that some of them do make serious attempts at structuring their efforts. There are, for example, many observable consistencies from classroom to classroom in the kindergarten program. However, generating a comprehensive set of objectives and activities is a difficult task; and it must be followed up by inservice which effectively promotes adherence to objectives and activities by program personnel--also a difficult task. In addition, some of the Title I program types--e.g., counseling--are particularly resistant to structure. In many cases, lack of structure in Title I programs is not due to negligence by program coordinators and personnel; some of them are working hard to overcome the difficulties inherent in attempting better organization of their efforts.

Design revisions. There will be two major thrusts to our design for future evaluation of Title I programs in Albuquerque, one involving product evaluation and the other process evaluation.

Our proposed product evaluation can best be characterized as an inductive search for effective variables. Although Title I programming overall does not seem to systematically raise children's reading gains, some of the children served by Title I do

make substantial gains. If we can identify the variables or combinations of variables that account for their gain, then Title I programming in Albuquerque can move in the direction of systematically manipulating these variables for all children in the programs.

Our primary approach to this problem will be statistical. Analysis to date has suggested that teacher and grade variables account for some of the variance in reading gain. We hope to look at the relationships between a number of variables and reading gain: children's grade, ethnic group, sex, entry score, school, teacher (both Title I and regular classroom), need for various services, and extent of services received; teachers' ethnic group and sex; and particular combinations of these variables--e.g., same sex and ethnic group for student and teacher as opposed to different sex and ethnic group. In addition, we will break out children with good reading gains and check the values of the above variables for this group. Do the successful children tend to have particular teachers, or combinations of teachers; do they tend to have higher entry scores; do they tend to be older, or female, or both, or neither?

An impressive amount of recent research indicates that the critical variable in school success is precisely that variable over which the schools have no control: the child's home background, especially his parent's socio-economic status. We anticipate uncovering some of the same kinds of uninfluenceable variables. Certainly, even in terms of our current information that greater gains are made by intermediate than primary children, Title I personnel may not be able to change that fact; but it will help to know it. If the facts cannot be changed, wisdom may lie in concentrating services where they will have some impact.

Our proposed changes in process evaluation will be much more difficult to implement, because they require a great deal of structuring from program coordinators. We hope that ultimately each Title I program, whether instructional or supportive, will define its objectives in sequential and measurable terms. The relationship of

the program's objectives to the overall objective of reading improvement should also be specified (with the exception of kindergarten). Ideally, each program should enumerate sequences of activities which lead to each objective. In some programs, diagnosis and prescription will also have to be built in much more specifically than they are now: criteria for including a child in a program and for determining the amount of service he receives will have to be spelled out and consistently applied.

Based on this information from program coordinators, we will construct process evaluation instruments which provide for systematic observation of program activities. The instruments will help assess an instructor's adherence to diagnostic criteria and to the sequence of activities that has been designed to promote program objectives, as well as the success of these activities.

At the end of the program year, we will measure attainment of each program's objectives and relate that attainment to reading gain (or for kindergarten, to school readiness). If the program is not attaining its objectives, some change in program activities will be indicated; the product and process evaluations should help in determining what changes need to be made. If the program is attaining its objectives without, however, affecting reading scores, change in the objectives themselves will be indicated. The consummation most devoutly to be wished, of course, is that each program meets its objectives and that there is a good correlation between program objectives and reading gain.

We have, at this point, an evaluation design which seems promising: each Title I program will measurably define its objectives, provide a curriculum, and specify criteria for serving children. Children who are served will be pre- and post-tested in terms of the objectives of the programs in which they are participating, as well as the overall objective of reading improvement. The Stanford tests will continue to be used to assess progress toward the overall objectives; instruments used to

screen children for program participation will probably serve in some cases for assessing attainment of program objectives as well.

Process evaluators will schedule visits to all program personnel while they are working with children. Screening, adherence to curriculum, and success of activities will be monitored. Program personnel will receive feedback on a regular basis throughout the year so that they can begin to correct problems as soon as they become apparent.

In each program except kindergarten, the end-of-year progress of served children toward individual program objectives and toward the reading objective will be compared with matched controls. Children in the control groups will be matched with treatment children for entry scores on the programs' own pre-tests as well as the Stanford reading subtests. (Some matching will be done by actually selecting matched controls; in other cases, variables will be controlled for statistically or even included as independent variables.) Within treatment groups, achievement of objectives will be compared across a number of possibly effective dimensions: e.g., sex, grade, regular classroom teacher, and individual program personnel. After these analyses, program personnel will know whether they are reaching their objectives and whether they are contributing to the reading objective; and these analyses, in combination with process evaluation input, should help assess program strengths and weaknesses.

When individual program results have been tabulated, results for all children who have been served in any program will be combined for an overall analysis of reading gains. If we can solve our problem of controlling for the possibility of need as an intervening variable, we will assess the interaction effects of the various programs on reading improvement. Finally, we will break out those children whose gains are good and compare them with children whose gains are poor in the hope that some pattern of variable values will emerge. These results will also be shared with

program personnel.

We also intend to expand our follow-up studies of Title I children. Kindergarten classes will be followed through grade school: Stanford scores and teacher evaluations for children who have attended Title I kindergarten will be compared, at the end of each school year, with comparable data for non-kindergarten children in all the appropriate grades in Title I schools. Reading Master Lists will be checked for the representation of kindergarten vs. non-kindergarten children. Participants in the Title I elementary programs will also be followed: each year's Master List will be checked for repeaters. If programming is effective and if the number of children who can be served remains essentially the same from year to year (which is primarily a function of funding level), then SAT cut-off scores for determining a child's eligibility should go up until any child who is below grade level is being served.

Problems of Implementing our Evaluation Design. The evaluation design that we have just described is at this stage a proposal only: it is one thing to design an evaluation and quite another to implement that design. The implementation problems that we have experienced and anticipate seem to be shared by other evaluators, especially those in larger public systems: in fact, most of the problems relate, directly or indirectly, to the realities of functioning in a large, bureaucratic system. By discussing our implementation problems here, we hope to accomplish two things. 1.) It may be helpful to other evaluators if we articulate common problems and present some potential solutions. 2.) These are the problems that we feel most impair our functioning. The first step in solving them, we think, is to define them as problems, clearly and openly.

It is not our purpose here to suggest that either the evaluation staff, the Albuquerque Title I program, or the Albuquerque Public School system has more than its share of such problems. In fact, our interactions with evaluators around the

country indicate that not only do they have as many or more problems, but they are farther from solutions. We feel fortunate in working in a system that permits open discussion of difficulties; again, we are convinced that only a system which openly confronts its problems has the potential for solving them.

One of our most comprehensive difficulties involves understanding and acceptance by school and program personnel of our role as evaluators. Although evaluation has been a widely publicized topic in recent years, it still seems to strike most program personnel as something of an imposition and an afterthought: evaluators are people who appear at the end of the year and refuse to see things as program personnel know them to be. Evaluators must be included in program planning; programs cannot be systematically evaluated unless they are designed with that purpose in mind.

The two most specific and troublesome problems are persuading program personnel to specify measurable objectives and to permit adequate controls. Most program personnel seem to consider such things as attitude change, happiness, and successful adjustment to school and life to be among their critical objectives for children. But they do not define those objectives in terms of observable behavior in any way that they themselves find acceptable. The objectives that can be acceptably and measurably defined--e.g., test performance, first-grade teachers' responses to children, and school attendance--are considered to be almost peripheral. Evaluating a program with reference to these objectives becomes meaningless: whatever the results, program personnel can continue to believe what they want to believe about the "real" program and the "important" objectives.

We are not unsympathetic to this point of view; attitude may, in fact, be a critical variable. But it seems pointless to devote all one's effort to such an invisible goal. The effort may be accomplishing everything or nothing--but no one will ever know for sure. Actually accomplishing a more limited objective seems

infinitely preferable to not knowing whether one has or has not accomplished a more comprehensive one.

We are also sympathetic to the dilemma that program personnel face with reference to adequate controls for evaluation purposes. These people are honestly concerned with helping individual children; they want to reach as many children as quickly as possible. The suggestion that some needy children be refused service so that their progress can be compared to the progress of program children is met with dismay and sometimes refusal. Again, it seems to us infinitely preferable to actually help half as many children than to not know whether or not one has helped all children, especially since documented successes can then be extended to all children.

The scope of Albuquerque's Title I programming creates another class of problems. With 350 regular classroom teachers administering Stanford tests to the 10,000 children in our twenty-seven Title I schools, it is extremely difficult to insure consistent and uniform administration procedures. Part of the problem is the teachers' rejection of standardized testing. It is almost as if they refuse to take standardized testing seriously. They claim that the tests' content is not related to their curricula; that the norming sample is different from their students; and that the tests do not elicit the levels of performance of which their students are capable. It is, in short, a waste of their time. With teachers making comments like one we overheard--"just finish this page, kids, and you'll never have to look at this booklet again"--we have to question the reliability of test results.

We have never suggested that Stanford results make a definitive statement of any kind about any child or any curriculum. We have, rather, always tried to keep standardized testing in what we consider to be its proper perspective: it provides a measure of how particular groups are performing relative to the national standard, and it predicts success or failure in regular school programs. In discussions with

Title I school staff, our instruments seem to be performing these functions fairly reliably. We suspect that teachers continue to reject the tests for exactly those reasons. Stanford scores in their schools are consistently low; they do not wish to be reminded that despite their efforts their students' school performance is, and will almost certainly continue to be, poor.

Another problem, definitely related to the public nature of our system, is that some of the most relevant variables to our analyses are taboo. The individual teacher variable looms largest in this context. Since there is only one staff member in each school for each Title I program (with the exception of kindergarten), we have been able to analyze the individual teacher variable for Title I programs without calling attention to the fact that we are doing so: we simply compare results across schools. And we have found that individual teachers account for more of the variance in children's performance than do any other variables. We suspect that what goes on in the regular classroom is, in turn, more influential than any Title I variable, because the children spend at least five hours in their regular classrooms for every hour they spend in special programs. We would like, therefore, to extend our analyses to the regular classroom teachers of Title I children. If we compare gains of children who have different combinations of Title I and classroom teachers, we will probably be able to account for more of the variance than with any other variable or combination of variables.

But everytime and everywhere that we broach this subject, we meet cries of dismay and prophecies of disaster. Our single actual attempt to openly include teacher variables in program evaluation bore out the prophecies. Even though we guaranteed anonymity for the individuals involved, the proposal generated a level of alarm that has still not been entirely resolved.

Part of the reason for this taboo is the strength and protectiveness of teachers' organizations. Certainly, those who work in the politically sensitive area of public education need protection from arbitrary or discriminatory sanctions. But once again, this seems to us to be a case of making important educational variables unnecessarily invisible and thus unmanageable. It is intuitively obvious that some teachers are better than others. If this fact is not openly admitted, it cannot be dealt with and there is no hope for improvement: the strong teachers will continue to be effective and the weak ones to be ineffective. If, on the other hand, the problem is brought out into the open, steps can be taken to upgrade the not-so-good teachers.

The public-school tenure system is also relevant in this context. As is typical in bureaucratic systems, employees are offered security. This fact causes endless difficulties. First and most obviously, programs can be permanently saddled with ineffective personnel who remain unchanged despite considerable effort to improve their performance. Second, and less obvious, those with the responsibility for hiring become gun-shy: they know they may be stuck forever with whomever they get. We have even witnessed the near-collapse of a good pilot program because funds were only available for a year: no one knew how to cope with the problem of one-year contracts for pilot personnel.

Perhaps the problem which causes us the most trouble on an almost daily basis is data processing. With the number of Title I children and programs in Albuquerque and the kinds of analyses we are attempting, we have gone well beyond the point at which desk calculators will serve. APS has a data-processing department and its own Honeywell computer; however, the department, like most in public school systems, has traditionally been geared strictly to business and accounting, not to instruction or instructional evaluation. All of the programs run on the system's machine are written by the local programming staff; none of these programmers, of course, was selected

for his statistical expertise. Even the program language used by the system is an accounting- rather than statistics-oriented language. Just creating a data file and programs for our analyses would require a year of a programmer's time; we have, in fact, been allotted one-eighth of a programmer's time.

Data-processing personnel are sympathetic to our dilemma; on many occasions they have gone out of their way to fulfill our requests for service. But they cannot change the fact that they have more programming to do than they have programmers available to do it. It is an understatement to say that, at this point, data analysis is a slow and frustrating process for us. We have explored the possibility of contracting with an outside agency for data processing services; our multiple regression analyses were performed by a private firm. Since data processing and analysis are part of our daily, ongoing business, however, we are convinced that this alternative would be expensive in terms of both money and time. Our proposed solution is to buy computer time from the University of New Mexico and set up our own data processing, using packaged programs for descriptive statistics, univariate analyses, and multivariate analyses. But a number of approvals are required before we can take this step. We are waiting. Meanwhile, last year's data sit largely unanalyzed, and this year's data will be crying for analysis in a matter of months.

The final and perhaps most critical problem area for us is dissemination. Title I evaluation in Albuquerque is producing a great deal of information with important implications for local Title I programming, for Albuquerque Public Schools as a whole, and even for federal programming as a whole. But there is simply no one to tell it to. Title I at the State level is primarily interested in our mean gains, which must be thrown in with the mean gains for every other Title I program in New Mexico to produce a single State-wide figure to send to Washington. Washington, presumably, is only interested in the State-wide figure as an input to a nation-wide

figure. (The violations to original data that are inherent in this kind of lumping process must make statisticians shudder.) A structure for disseminating research and evaluation findings has not been built in to either the Albuquerque Public School system or to Albuquerque Title I. We rely, at this point, on our personal contacts and persuasiveness to make our evaluation findings effective.

Accountability is a companion term to evaluation. The notion of evaluation for accountability is a prominent one in contemporary U.S. education--and one to which we subscribe. But accountability is impossible without structure: a program structure which permits systematic evaluation; an evaluation structure which provides dependable and useful results; and a dissemination structure which makes those results effective. We, as the evaluation component, are primarily identified with accountability, and we are attempting to create an evaluation structure which promotes real accountability. But, as we concluded from last year's evaluation, we cannot offer constructive help unless program personnel structure their efforts so that they can be systematically evaluated; nor can we effect implementation of our findings. The responsibility for supporting evaluation and for insuring that evaluation findings are effective resides above us in the Albuquerque and federal systems.