IPLE LINEAR REGRESSION VIEWPOINTS
ME 16, NUMBER 1, SUMMER 1988

Regression and Model C for Evaluation

Gail Smith, Keith McNeil and Napoleon Mitchell
Dallas Independent School District
Dallas, Texas

OVERVIEW OF SYMPOSIUM

The objectives of this symposium are to:

1) Provide a rationale for using regression analysis (specifically Model C) to evaluate educational programs.
2) Provide one example of an extensive Model C evaluation report.
3) Discuss assumptions of Model C and ways to deal with those assumptions.
4) Share examples of disseminating Model C results to decision makers.
5) Identify and resolve additional technical issues that evaluators need to be concerned about when implementing Model C.

We ask you to pretend that this is the Dallas Independent School District Board meeting. The program manager and evaluator are presenting the end of year evaluation results for a state-funded compensatory education program. The evaluator has briefed the program manager and the report was delivered to the school board approximately two weeks ago. We must assume, though, that no members thoroughly understand the report, mainly because most have not read it in anticipation of being briefed.

The presentation will be made by two evaluators from DISD. Gail Smith will be playing the role of program manager in presenting the basic program. Keith McNeil will be playing the role of evaluator in presenting the evaluation results. A third evaluator from DISD, Napoleon Mitchell, will be playing the role of court-appointed auditor, questioning the precedures, results, and interpretations. (Those of you who do not have the pleasure of working under the constraints of a court order may want to think of Napoleon as a board member who has a Ph.D. in statistics and doesn't mind you knowing it.) We would appreciate you asking your questions only after the auditor is satisfied that all his questions have been asked/answered. The last ten minutes of the symposium is reserved for the comments from our distinguished discussant, Dr. George Powell of the Educational Testing Service.

DESCRIPTION OF TREATMENT PROGRAM

The goal of the Reading Improvement Program was to narrow the gap in reading performance between lower and higher achieving students as well as minority and White students in the District. Objectives for accomplishing this goal included: a) providing an additional two-semester, reading course with a restricted teacher pupil ratio of 1:20, b) providing special curriculum materials in logic, vocabulary, comprehension, and study skills, and c) providing staff development on effective instructional strategies in reading to participating teachers. The additional language arts course, focusing on reading, was required for students in grades seven and eight who scored below the 40th percentile in Reading Comprehension on the Iowa Tests of Basic Skills (ITBS). All students scoring below the 40th percentile at all 24 District Middle Schools were eligible for the program with two exceptions. Students in special education classes and students in the two beginning levels of English-as-a-Second-Language classes were not eligible.

Characteristics of students enrolled in the program are presented in Tables 1 and 2. The figures in Table 1 indicate that nearly half the

Table 1

Number of Students Enrolled
and Not Enrolled in RI Course
Fall, 1984

Enrolled in RI Course	Grade 7	Grade 8
Yes	4285	4790
No	5374	4383
Total	9659	9173
% of Total in RI Course	44	52

students in the District middle schools were enrolled in the program in the fall of the second year. The analysis of program effectiveness was conducted using ITBS reading comprehension test scores for both Spring 1984 (pretest) and Spring 1985 (posttest). The number of students represented in this analysis is provided by race and grade in Table 2. Ethnic minority students comprised 87% of the total number of participating students at both grades seven and eight.

Table 2

Grade	Stat	Ethnicity Black	Hispanic	Asian/Indian	White	Total
7	N	2111	591	36	398	3136
	%	67.3	18.8	1.1	12.7	
8	N	2580	705	20	482	3787
	%	68.1	18.6	0.5	12.7	

Since the districtwide percentage of minority students was 76%, the RI program was focusing on minority students.

IMPLEMENTATION FINDINGS

The RI program in grades seven and eight was implemented much better than last year, though there were improvements needed. The lack of a program manager with clear lines of authority resulted in lack of communication and slow or erroneous implementation. Staff development sessions were less than successful because of redundancy of topics and timing of material.

Almost all of the classrooms observed appeared to be conducive to learning, although some did have an enrollment of more than the maximum of 20 allowed by the guidelines. Teachers were using the RI texts and support materials, but few were using teaching techniques considered beneficial for these kinds of students.

Few interactions were initiated by students with the teacher controlling the interactions. Although most teachers provided positive reinforcement, not all teachers provided at least five instances of positive reinforcement. The instructional climate was judged to be better in the RI classes than in the regular language arts classes, both in terms of how well the instructional time was used and whether the instruction was conducive to learning.

ACHIEVEMENT FINDINGS

Results for Grade Seven. A total of 3135 RI students had both pre and post scores, although the scores of 151 of these students were eliminated because their post score was considered too deviant in respect

to gains which were either too high or too low to meet normal expectations. Students in RI gained from 30.1 to 32.3 NCE units. But since RI students were selected into the program according to their pretest scores, we would expect the regression effect to raise their scores. RI students also gained more than the comparison group whose pretest scores were above the 40th percentile (2.3 mean gain vs. -5.9 mean gain for the comparison group). Again, though, the regression effect would have predicted the general trend of these results, i.e. the initially higher scoring comparison group showed mean losses while the initially lower scoring RI students showed mean gains.

A significant second degree fit to the data was discovered in the seventh grade comparison group. Hence the Model C analysis employed a second degree curved line of best fit. the curved line of best fit was the same for both the comparison group and the RI group, hence for these eighth grade students there was no effect due to participation in the RI program (See Figure 1).

Results for Grade Eight. A total of 3787 RI students had both pre and post scores, although the scores of 184 of these students were eliminated because their post score was considered too deviant in terms of expected gains or losses. RI students gained from 30.2 to 34.8 NCE units. But since RI students were selected into the program according to their pretest scores, we would expect the regression effect to raise their scores. The RI students gained more than the above 40th percentile comparison students, but again the regression effect would have predicted this outcome.

There was no second degree curvilinear fit found in the eighth grade data, so only linear trends were investigated. Since a significant

interaction was found, an overall program effect was not investigated. The analysis was concluded with the findings of a significant apptitude-treatment interaction. The lines of best fit for the eighth grade are depicted in Figure 2. The RI program is most effective for those students who have the lowest pretest scores. Those students at the program cutoff gain very little from the extra RI class.

ALTERNATIVE EVALUATION MODELS

There are three major ways we could have evaluated this program. These three ways were documented and described by Tahorst, Lmadge and Wood in 1975.

First, we could have compared the performance of DISD children with what we would expect them to do if they were like the national norm group. This has been referred to as the Model A approach, wherein we use the pretest achievement level as the expectation for the posttest performance.

Two major assumptions in the use of this model cannot be met. The selection of students into the program should be independent of the pretest score, otherwise simple regression to the mean can account for substantial movement to the total group's mean. This was the situation in the RI program, as the pretest measure also served as selection into the program.

The second assumption of Model A which cannot be verified is that the students in the norming sample who are at the same pretest percentile levels are like those being evaluated -- like in the sense of demographics and in terms of quality of regular educational curricula. We know that most of the DISD students are inner city students, with a high concentration of low SES students. Therefore, we can't assume that our students are like the national norming sample. The test that we use does have large city norms. Although DISD students consistently score high, we cannot determine if our students gain more than other large city students. The high scores may only reflect higher initial achievement levels of our students. That is, the question of the quality of a program demands assessment of student growth.

Second, we could have used was a local comparison group to evaluate the RI program. This type of evaluation is referred to as Model B in the literature. Model B is difficult to implement in most educational settings, as in this one, because the model requires that some students (who are otherwise qualified) not receive the special treatment. All students scoring below the cutoff of the 40th percentile were supposed to receive the treatment, thus leaving no students for the comparison group. What actually happened was some students below the 40th percentile did not receive the RI course. Some of these students were in special education classes and some received the RI course only one semester. The remaining students did not receive RI for undocumented reasons. It was our educational guess that many of these students were not enrolled in the RI course for educational reasons which would indicate a higher posttest level than indicated by their pretest (e.g. student is really a high achiever, she just didn't pretest well).

The third and final model utilizes a local comparison group which is acknowledgely different at pretest time. The model capitalizes on the fact that this local comparison group receives the same regular curriculum. The expected posttest performance of the treatment group (RI students) is estimated from the performance of the comparison group. This model assumes that the achievement gain is consistent across pretest levels. One of the major problems of Model C is the determination of this consistent trend in achievement gain. Is the trend linear or of some other nature? Another problem is that the presence of erroneous outliers can unduly affect calculations of this trend. Outliers do not affect the calculations of statistics in other models as much as in Model C.

Exhibit 1. Summary of Models

Model Name	Comparison	Expected Post Performance	Problems	Advantages
Model A	students as their own comparison	pretest level	1. selection on pretest 2. students in norming sample—ethnicity, size, quality of program	1. easy to compute by hand 2. similar to what was done in past
Model B	local students who do not receive treatment	posttest of comparison students	1. students denied service 2. requires testing of additional students	1. both groups of students receive similar regular curriculum
Model C	local students who do not receive treatment	predicted from comparison students	1. linearity 2. outliers 3. calculation and interpretation	1. both groups of students receive similar regular curriculum 2. don't have to deny services to some students 3. can test for aptitude by treatment interaction 4. can reflect non-linear reality

Model C was chosen as the best model to evaluate the program because students were selected into the program on the basis of their pretest scores, and most students below the cutoff score were served. Those that were not served did not constitute a valid comparison group as many were suspected to have been exempted because their pretest score was felt to be not indicative of their true performance.

CONCEPTUALIZATION OF MODEL C

Whether or not the RI scores are elevated is the first question to be answered. We can begin to conceptualize the model by looking at Figure 3. All those students who have a pretest score below 40 are placed in the RI program as well as the regular curriculum, while all those who have a score of 40 and above are not allowed in the extra RI course and, hence, only receive the regular curriculum. After eight months of instruction, the posttest scores are obtained. The straight line of best fit is calculated for the comparison group. This line indicates the expected posttest performance for students at each pretest score. (See Figure 4.) If the line fits well, (correlation above .4) then we can proceed and assume that the straight line can be extended down into the range of scores of the treatment group which received RI. (See Figure 5.) We know, though, that the students below the cutoff not only received the regular curriculum but also received the RI curriculum. Therefore, the posttest scores of those receiving RI should be higher than if they would not have received RI. (See Figure 6.)

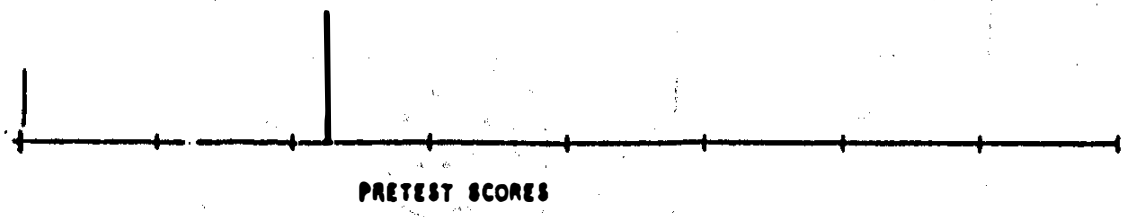

Figure 3. Selection of students into program, based on pretest score.

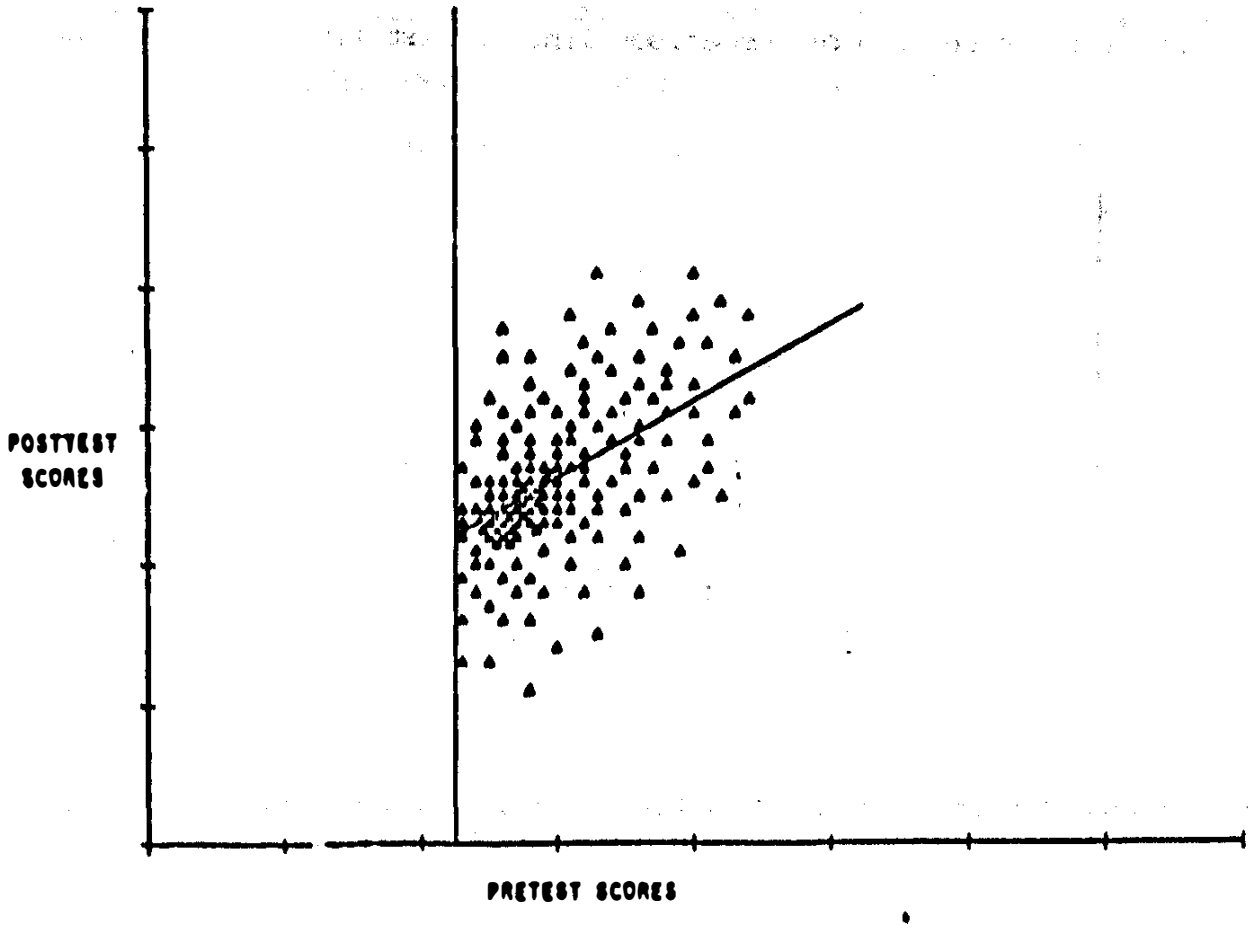

Figure 4. Line of best fit in comparison group.

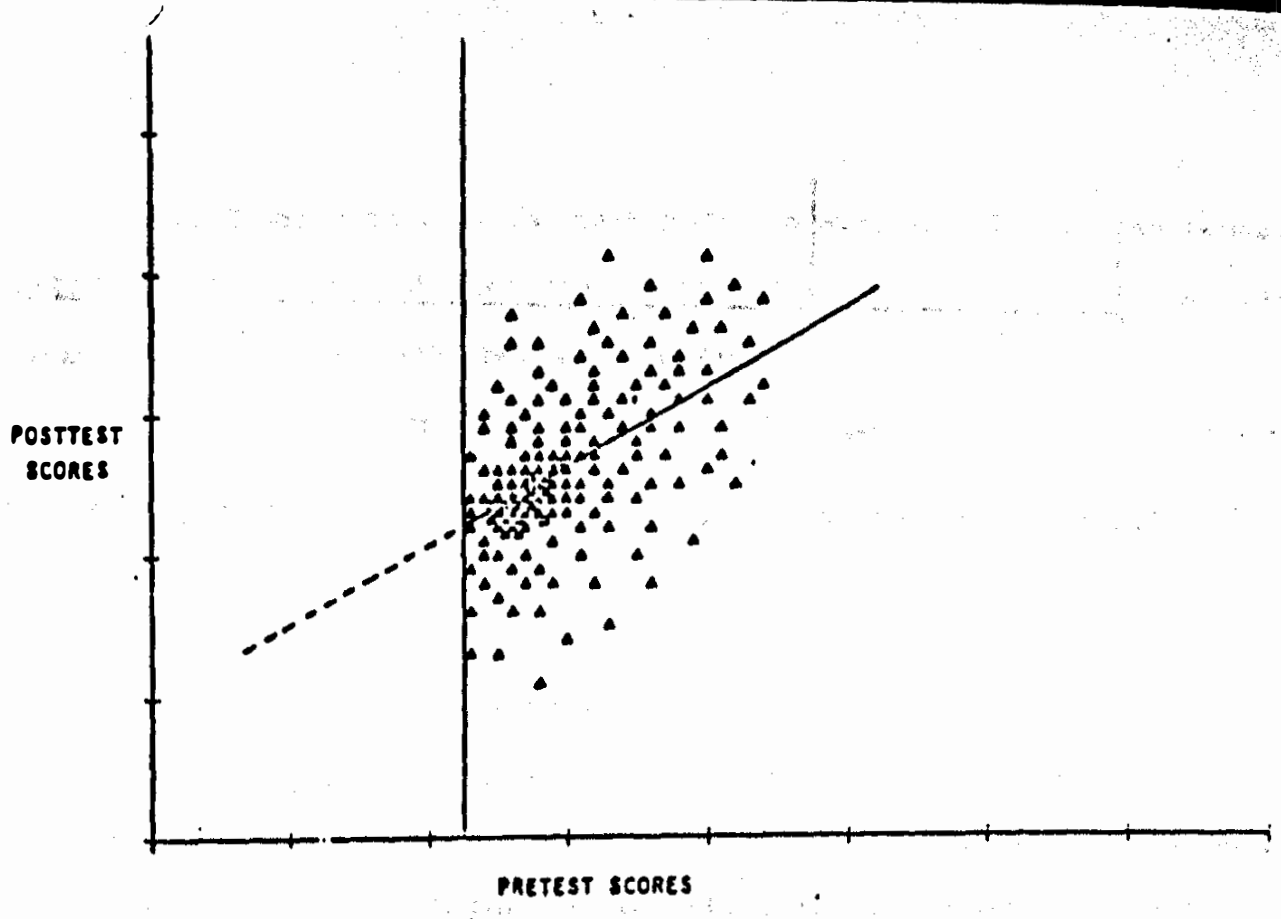

Figure 5. Extension of comparison line of best fit into treatment group.

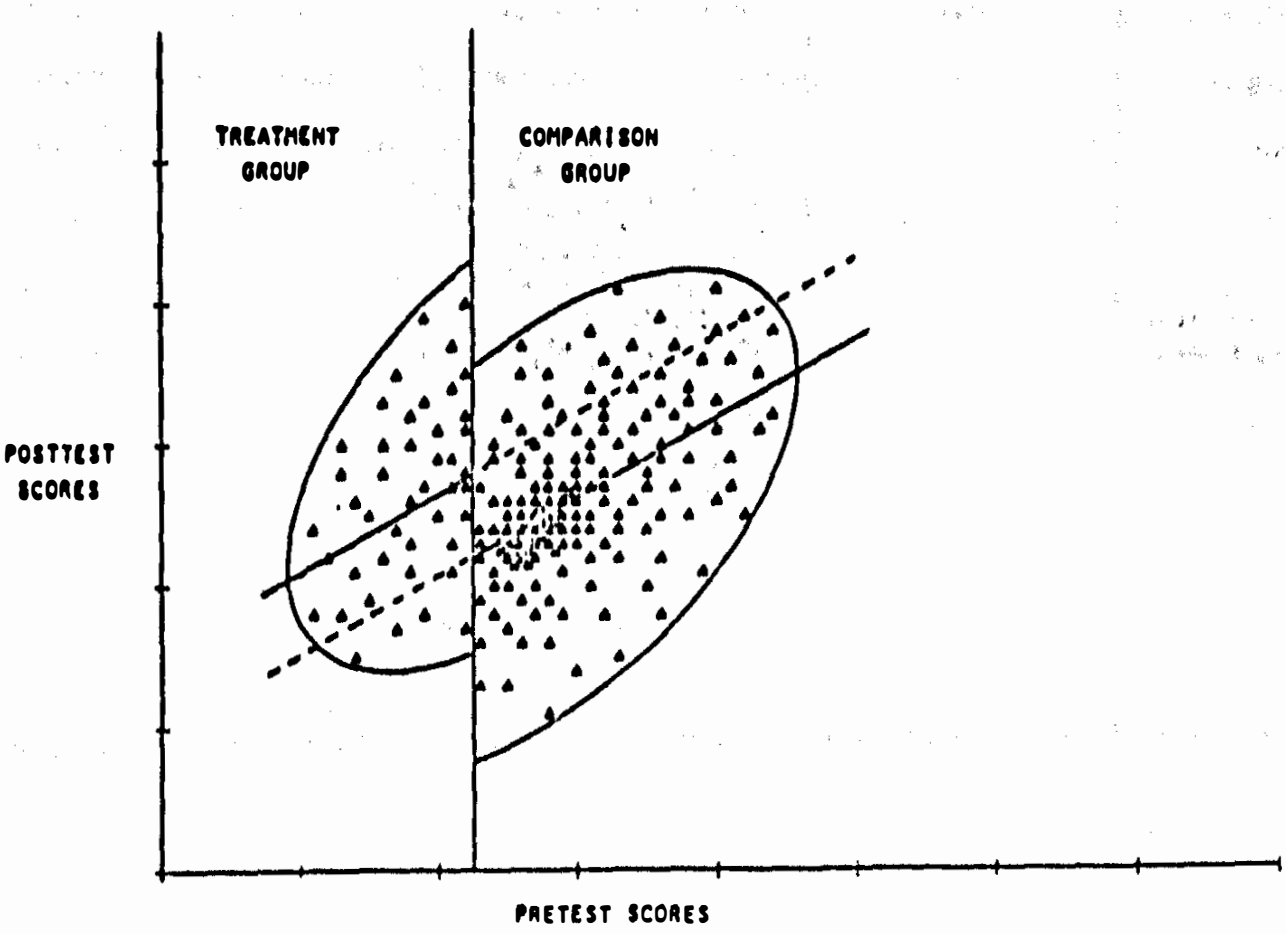

Figure 6. Model C illustration of treatment effect.

A second question of interest is whether the elevated effect was consistent across pretest scores. It might be that the RI program is especially effective in producing higher than expected gains for the lowest achieving students. (See Figure 7.) Or, the RI program may be especially effective for the highest students in the treatment. (See Figure 8.) Different program recommendations would, of course, result from these two different findings (findings which, by the way, would not surface in a Model A or Model B analysis). Thus, the second question of interest is, "Is the RI treatment differentially effective over the various pretest levels?" Another way to verbalize this interaction question is, "Is the RI line of best fit parallel to (exhibit the same slope as) the line of best fit for the comparison group?"

Model C, as any statistical question, can be tested with the general linear model. The full model contains all the information identified in the question (research hypothesis). Restrictions (identified in the question) are made on the full model, resulting in the restricted model. the difference in the number of pieces of information in the full and restricted models is equal to the number of restrictions. The general F-test formula is:

$$F = \frac{(R^2_{FULL} - R^2_{REST}) / (pieces_{FULL} - pieces_{REST})}{(1 - R^2_{FULL}) / (N - pieces_{FULL})}$$

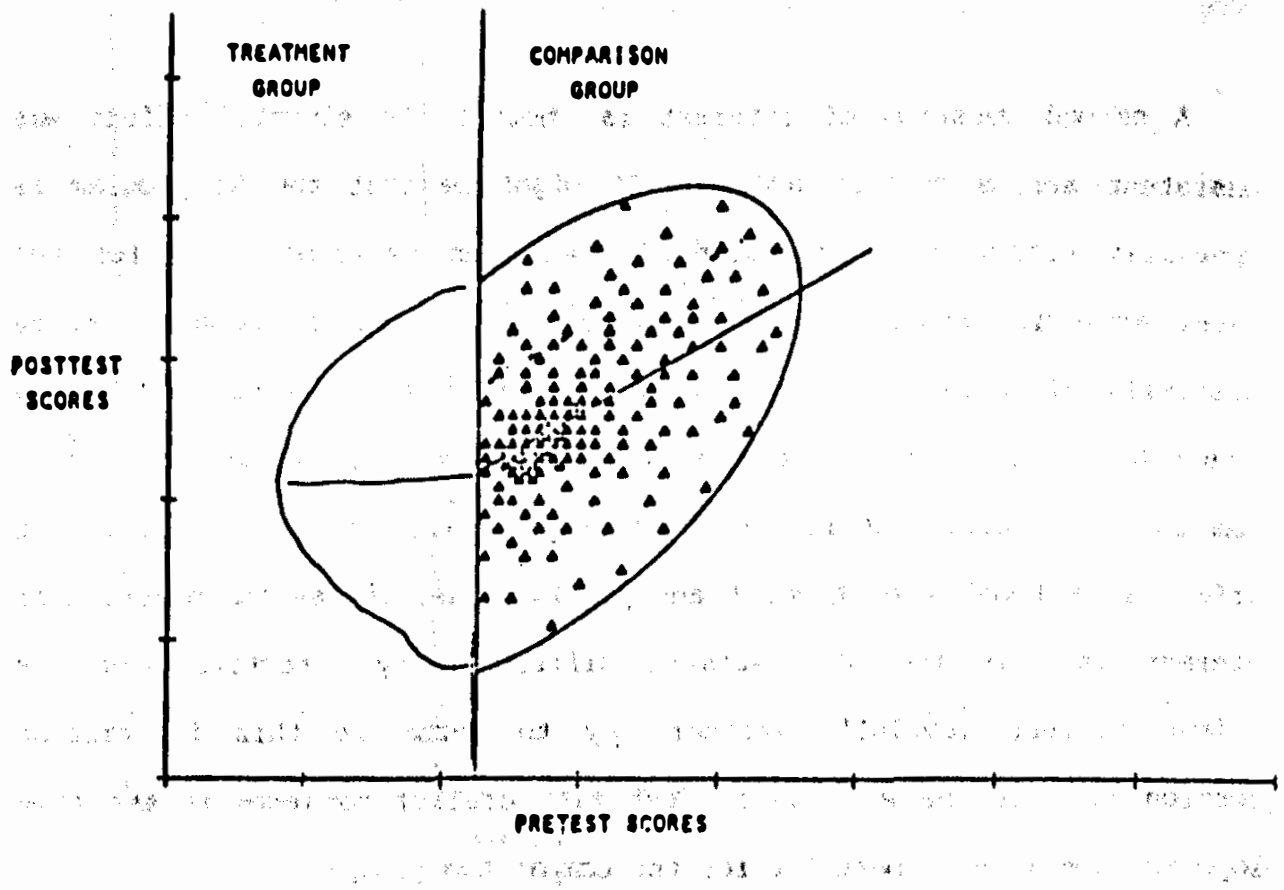

Figure 7. Model C illustration of treatment especially effective for l achieving treatment students.

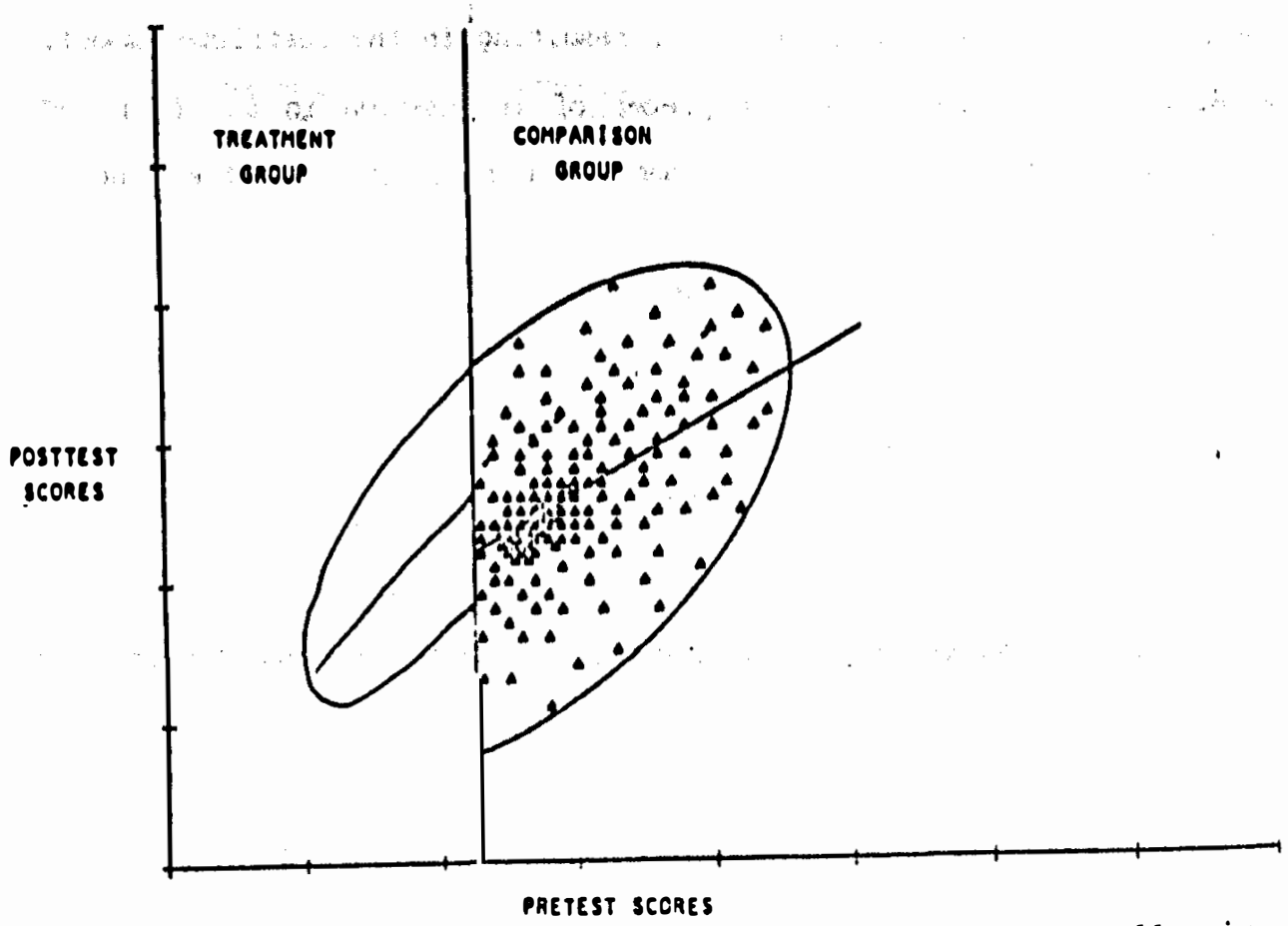

Figure 8. Model C illustration of treatment especially effective f high achieving treatment students.

MISCELLANEOUS DATA ANALYSIS TOPICS

Scales

All test information was transferred from percentiles to NCEs. NCEs are Normal Curve Equivalences which are a normal distribution transformation of percentiles. NCEs are an equal interval scale, therefore amenable to statistical manipulation. They have a mean of 50 and a standard deviation of 21.06.

Comparison groups

The comparison group should be receiving the regular curriculum received by the treatment group. In Dallas, most of the students above the 80th percentile enroll in an honors English course. Therefore, students above the 80th percentile were excluded from the analyses. Some students who should have been in the RI program because they had a qualifying pretest score below 40 were not given the special treatment. Before these students were combined with the regular comparison group, they were analyzed to see if they functioned differently.

Outliers

Students whose posttest scores were more than two standard errors of estimate beyond their predicted posttest score were eliminated from the analyses. The statistics used for a given student came from that student's group, RI comparison above 40, or comparison below 40.