DOCUMENT RESUME

ED 273 674 TM 860 526

AUTHOR van der Ploeg, Arie J.; And Others
TITLE An Investigation into the Relative Effectiveness of Remediation Programs.
PUB DATE 16 Apr 86
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Core Curriculum; *Effect Size; Elementary Education; Evaluation Methods; *Evaluation Problems; Meta Analysis; Norm Referenced Tests; *Pretests Posttests; Program Effectiveness; *Program Evaluation; Regression (Statistics); *Remedial Programs; Testing Problems; Test Validity
IDENTIFIERS *Education Consolidation Improvement Act Chapter 1; Normal Curve Equivalent Scores

ABSTRACT

The effectiveness of an Education Consolidation Improvement Act Chapter 1 program was studied in a large urban school district. A common problem in evaluating remedial programs is that pretest/posttest achievement gains often indicate the effects of the entire curriculum, rather than those specific to the remedial class. In this district, over 50,000 Chapter 1 students were enrolled in more than 250 elementary schools. A conceptually simple method was needed for studying the effects of the remedial program versus the core curriculum. School regressions, residuals, mean residuals for Chapter 1 students at each school, and a number of mean residuals were computed from a database containing pretest and posttest scores. Residuals from within-school regressions were used to model the effect of the entire curriculum. Each school's mean residual indexed the effectiveness of its own Chapter 1 program. Results indicated one-third of the school's mean residuals for Chapter 1 students were positive. A large correlation matrix of variables which might explain the variance patterns was not conclusive. Only a few variables were modestly correlated with the residuals. In addition, few of the explanatory variables correlated with each other as expected. Measuring teacher characteristics might have been helpful in this project. (GDC)

AN INVESTIGATION INTO THE RELATIVE EFFECTIVENESS OF REMEDIATION PROGRAMS

Arie J. van der Ploeg
Linda K. Junker
Robert E. Bole
William K. Rice

Department of Research and Evaluation
Chicago Public Schools

Presented at the annual meeting of the
American Educational Research Association
April 16, 1986
San Francisco, CA.

AN INVESTIGATION INTO THE RELATIVE EFFECTIVENESS OF REMEDIATION PROGRAMS

Arie J. van der Ploeg, Linda K. Junker, Robert E. Bole, and William K. Rice
Department of Research and Evaluation, Chicago Public Schools

Most remediation programs enroll a student for only one school period each day. Most of the student's school day is spent not in remediation but in the regular classroom working on the same materials as her/his peers not in the remediation program. This fact has several implications for evaluation of such projects, implications largely ignored in the literature. A prime example is the study of the effectiveness of ECIA Chapter 1 (formerly ESEA Title I) programs.

The Problem

The typical evaluation of a Chapter 1 project computes the mean pre- and posttest Normal Curve Equivalent (NCE) scores for students in the project and attributes success to the project if the posttest mean exceeds the pretest mean. But, students' learning is acquired and cumulates over the entire school day and throughout the school year. The fact that the posttest mean on an assessment measure given to participants in a remedial project exceeds the pretest mean is, at best, only indirect evidence of any effect due to the project. More likely, the learning students gain outside the project is being measured, since students spend much more time each day of the school year working in the regular, not the remedial, classroom.

This appears even more true given that most remedial programs, particularly ECIA Chapter 1 programs, are commonly assessed by means of nationally normed standardized achievement batteries, instruments which are intentionally designed to be sensitive to the whole of what is learned in school. It is rare to encounter a Chapter 1 evaluation using an assessment instrument precisely calibrated to the skills being developed by the Chapter 1 project [see Carter, 1984, or Gabriel, 1985].

Large scale evaluations of educational programs have often found greater variation between sites within programs than between programs [e.g. Coleman, Campbell, Hobson, et al., 1966]. This should not surprise, since variations in the effectiveness of the general program of instruction are probably greater than any variation created by programs taking up only a small portion of the instructional day.

The foregoing suggests that evaluation of remedial programs should focus either on outcomes unique to the remediation and not taught in the core curriculum, or on the value added by the remediation to the core curriculum's contribution to students' learning. Both these approaches are difficult to implement for most programs sponsored by Chapter 1 of ECIA. Unique outcomes are generally not desired nor measured since the goal is to provide students with the skills needed to function effectively in the regular classroom. Program designers therefore see general measures of learning to be appropriate criteria. However, the value added by a remedial program cannot be determined without knowledge of the contribution to learning of the core curriculum. This requires either random assignment to control and treatment groups, a procedure considered anathema by most school districts, or the presence of a naturally occurring control group, an unlikely event in Chapter 1 since all eligible students are typically assigned to the program.

The Search for Solutions

This paper reports on procedures investigated in a large urban school district to evaluate the effectiveness of its Chapter 1 program, controlled for the effect of the core curriculum. The district's Chapter 1 project annually enrolls over 50,000 students at more than 250 public elementary schools. Each eligible school designs its own Chapter 1 project, subject to central approval.

The schools participating in the project differ considerably in their effectiveness in teaching the core subjects. Staff of the district's evaluation unit agreed that assessment of the system-wide effects of Chapter 1 must account or control for this variation. They also agreed that the technique to accomplish this should be conceptually simple; complex designs, it was feared, could intimidate or confuse the intended audience of administrators, board members, and the public. Hence, multivariate models relying on sophisticated covariation or partialling techniques were ruled out.

Contrasting the mean achievement of students in and not in Chapter 1 programs at each school and then analyzing the difference of these means across schools initially seemed a promising approach. The difference between the groups within each school could be conceptualized as the "value added" by the remedial program. Aggregating to the school level would permit the introduction of school level variables into the analysis without the problems attendant upon mixing school and student level data.

Constructing an appropriate index to measure the difference between the Chapter 1 and non-Chapter 1 means at each school was a major concern. The index should assess the effect of the core curriculum and be able to specify the additional effect due to the remediation. In addition, it should be sensitive to any differences in pretest scores both within and between schools. Indices derived from gain scores were first considered but discarded because they were theoretically unreliable and exhibited considerable collinearity in the more complex indices.

The techniques developed for meta-analysis were next considered. The usual definition of effect size,

$$g = (\overline{Y}_e - \overline{Y}_c)/S_c,$$

seemed inappropriate. The difference between the remediation program's participants' posttest (Y_e) mean and the posttest mean of the students not participating would generally be negative--Chapter 1 students typically score lower--and the effect of the core curriculum was not being isolated.

Use of residuals from within-school regressions appeared to offer a solution. Regressing all students' posttest scores on their pretest scores within each school would model the effect of the entire curriculum. The mean residuals of the Chapter 1 participants could then be collected from the within-school regressions school by school. Each school's mean residual would index the effectiveness of its own Chapter 1 program, since the additional instruction the Chapter 1 program provided--instruction not received by those not participating--would augment the value of the residuals for the Chapter 1 students. Standardized and unstandardized residuals would each add to the interpretation: standardized residuals could be safely aggregated across schools; unstandardized residuals would provide an index, expressed in the metric of the original measure, of the absolute amount of progress attributable to the remediation.

Cautions

This approach, it was realized, would necessarily be conservative. The within-school regressions model the joint effect of the core curriculum and the remedial program. If the number of students enrolled in the remediation were the minority, the effect of the core curriculum would dominate the regression. As the proportion of Chapter 1 students increased, the regression line would estimate the general school effect less accurately. This implies that positive residuals for the Chapter 1 students should be interpreted to confirm a strong effect for the program whereas negative or zero residuals should be interpreted as denying a strong effect but not ruling out the presence of an effect.

It would have been preferable to exclude the Chapter 1 students from the within-school regressions. The regression line would then not be contaminated with any effects due to Chapter 1 programs, giving a "clean" estimate of the effect of the school's regular instruction. The measure of a Chapter 1 effect would be obtained by calculating "residuals" for Chapter 1 students using the coefficients obtained from the regressions. However, excluding the lowest achieving students from the regressions altered the slopes of the regression lines significantly, thereby introducing new inaccuracies and uncertainties into the computed "residuals." It was decided to accept the bias of the approach previously outlined. To build an index measuring relative effectiveness was deemed more important than precise estimation of absolute progress. Furthermore, partialling or covarying on the proportion of students enrolled in Chapter 1 in each school should limit the contamination inherent in the chosen approach.

This approach has the virtue of being analogous to meta-analysis, if effect size is defined as the observed posttest mean for remediation program students less the mean of their expected posttest scores, i.e.,

$$g = (\overline{Y}_{obs} - \overline{Y}_{exp})/S_{exp}$$

where Y_{exp} is computed from the within-group regressions. The $\overline{Y}_{obs} - \overline{Y}_{exp}$ term is, of course, another expression of the mean residual for the group.

Y_{obs} is the observed, tested achievement of the student; Y_{exp} is the estimate of the contri-bution of the core curriculum. The difference is the amount of learning which may be attributed to the remediation.

Procedures:

A computer file containing Chapter 1 participation information already existed and was used to identify the pre- and posttest scores of Chapter 1 students on the annual citywide testing files. The procedures available in SAS (SPSS or a similar software package could just as easily have been used) greatly simplified the running of 250 individual regressions, the computation of residuals, the computation of mean residuals for Chapter 1 students at each school, and the creation of a school level file containing only the mean residuals.

Several different mean residuals were computed. Some Chapter 1 programs were pull-out programs, others used a reduced class-size self-contained model; some programs focused on reading instruction, others on mathematics, still others taught both. Mean residuals were computed at each school for each of these categories of Chapter 1 programs. If a category was not present at a school, a previously defined missing value was inserted in the record.

The next step consisted of creating school level files of variables to be used to explain the variation in the residuals between schools. A variety of data were collected from existing data bases, edited, and compiled. These included such variables as: student and teacher attendance rates, stability of the student body and teacher turnover, principal's age and experience, teachers' education and experience, the poverty index of each school, the mean pretest score for all students and all Chapter 1 students at each school, the percentage of students retained in grade, the school's total enrollment and the proportion served by Chapter 1, the total and per Chapter 1 pupil costs of each school's Chapter 1 project, the racial/ethnic composition of the school's student body, and the grades served by each school.

More complex variables such as the amount of time spent in Chapter 1, the student:teacher ratio adjusted for the time spent in Chapter 1, time-on-task, ratings of staff, program, and administrative quality, curricular content and method of the Chapter 1 project, classroom climate, staff morale, degree of program implementation, and so forth could not be created within the time limits of this study or were not available. Unfortunately, these omitted variables appear on their face to have a more direct bearing on the effectiveness of instruction than do many of the variables used. It was hoped that at least some of the variables which were used would serve as proxies for these omitted variables.

Results:

About one-third of the schools' mean residuals for Chapter 1 students were positive. Given the expected conservative bias of this procedure, this seemed appropriate--not that more positive residuals would not have been preferred. Inspection of the rank ordering of schools on each of the residuals cast no doubt on the integrity of the results. The variance of the mean residuals was very limited, which is congruent with almost 20 years of local evaluations of the Chapter 1 project: typically, small gains are recorded each year and very few schools display gains or losses greater than a few NCE units. The mean residuals for mathematics showed a slightly greater variance than the residuals for reading, an expected result given the greater discriminating power of mathematics tests generally. Table 1 below sets out some basic descriptive information for the reading and mathematics residuals calculated across all Chapter 1 students.

TABLE 1

DESCRIPTIVE STATISTICS FOR READING AND MATHEMATICS RESIDUALS

						Percent
	Number of					between
Residual	schools	Mean	S.D.	Minimum	Maximum	-0.5 and 0.5
Reading	238	-0.081	0.188	-0.808	0.355	96.2
Mathematics	197	-0.070	0.285	-1.302	0.892	92.4

A large correlation matrix of the residuals and the explanatory variables was next built in order to determine which variables would become candidates for models to explain the variance patterns. This produced an unexpected result: with few exceptions, none of the variables correlated with the residuals. Those which did correlate, correlated only modestly--which may be an overstatement. Even more surprisingly, few of the explanatory variables correlated with each other in the manner expected.

Inspection of the programs used to create these variables located no errors. Some deliberation, however, led to ad hoc reasons why many did not correlate well with each other. On reconsideration it became clear that the variables dealing with principals and teachers did not correlate highly among each other because the school district has explicitly followed a policy of equity in teacher and principal assignment with respect to such variables as race, experience, and training. The failure to find intercorrelations attests to the success of this policy. Pupil attendance at these elementary schools has essentially no variance due to mandatory enrollment policies. The Chapter 1 monies available to these schools are allocated based on the number of students with poor academic achievement and poverty backgrounds; hence, the amount available per participating student at each school varies little.

The failure to find correlations and the consequent inability to build explanatory models is a temporary setback. The test of the procedure must lie in better specification of the explanatory variables. Several of the more complex variables previously discussed are being created. When that work is completed, the utility of the procedure will be tested again.

Discussion:

The procedure described in this paper emphasizes the use of standardized test results to create an index of program effectiveness. It must be recognized, however, that outcomes other than test results are inherent in the schooling process [Dreeben, 1968]. The decision to focus on test results, in this study as elsewhere, should not preclude investigation of other outcomes.

Although the explanatory variables used in this study were admittedly inadequate, their failure to confirm the procedure does suggest reconsideration of the logic underlying the procedure. The fact that the regression line, intended to estimate the general effectiveness of each school's instruction, is influenced by the inclusion of Chapter 1 students is a weakness. However, short of a truly randomized design or witholding remedial funds, no alternative appears a better one. If the residuals are treated as an index and not as a statement of the actual effect size, and if any explanatory analysis adjusts for the proportion of each school's students enrolled in remediation, then the ranking based on the adjusted index should serve its primary purpose: to discriminate schools efficiently and accurately with respect to the effectiveness of their remedial programs.

Despite the ad hoc reasoning presented with respect to the, for our purposes, inadequacies of the explanatory variables used, it is possible to argue that they are adequate and the pattern of no significant correlations is the correct one. This implies that the effectiveness of these Chapter 1 programs do not correlate with these variables because the Chapter 1 programs do not add enough to what students learn. Their effect is limited, muted by measurement error and the difficulty of significantly improving the learning of students achieving much less than their peers. This is not inconsistent with local evaluations or with the long-term and short-term national Chapter 1 results discussed by Stonehill [1985]. Even those studies which report significant long-term benefits to participation in remediation programs [e.g. Berrueta-Clement, Schweinhart, Barnett, et al., 1984] report only minimal test score changes in the first years of exposure to the program.

Other factors not addressed in this study may also muddy the waters. Previous local research has made clear that variations in test results among grades within schools are not uncommon and sometimes large. Possibly the regressions should be done for each grade within each school. This complicates the analysis somewhat, but conducting it grade-by-grade within school is only an extension of the existing procedures.

As Carter [1984] points out, there may also be interactions between a remedial program and the students served: a program may appear "effective for students who were only moderately [educationally] disadvantaged, but . . . not improve the relative achievement of the most disadvantaged part of the school population."

Schools may be perceived as entities comprised of multiple wills and purposes: the student, the classroom teacher, the remedial teacher, the subject area specialist, the counselor, the principal, the district staff. Often, schools are less than the ideal unified, purposeful organization. "Loosely coupled" is the sociological jargon [March and Olsen, 1976]. Educational processes, organizational structure, and outcomes may become "disconnected" [Meyer, 1980] from each other. To expect consistent, definable, measurable effects of schools under such circumstances may be unreasonable. Taken further, this reasoning may imply that the appropriate level of analysis to locate program effects is the classroom, not the school.

More cogently, it may be that the most important element in the learning of Chapter 1 students is the enthusiasm, diligence, capacity, and ability of their teachers. Omitting variables assessing the teacher may permanently limit the meaningfulness of analyses conducted along the lines discussed here, or, for that matter, any analyses of educational effectiveness. Measuring these traits is however no simple task and fraught with procedural, social, and political difficulties.

Despite the complexity of this discussion, the procedure advocated is not complex. Bivariate regressions and computation of means are the only statistical tools required. The meaning of the regression lines and the residual is readily apparent to most audiences, if jargon is avoided. Treating the mean residuals as an index of relative effect and not absolute achievement also helps. Operationally, the procedure is also simple, since it can be implemented easily and quickly, assuming the data are available, on software packages such as SPSS or SAS. We urge other school districts to investigate this technique and report to us on their reactions to its adequacy and utility.

REFERENCES

Berrueta-Clement, J.R., Schweinhart, L.J., Barnett, W.S., Epstein, A.S., and Weikart, D.P. (1984). *Changed Lives: The Effects of the Perry Preschool Program on Youths through Age 19*. Ypsilanti: High/Scope Press.

Carter, L.F. (1984). The Sustaining Effects Study of Compensatory and Elementary Education. *Educational Researcher*, *13*, 4-13.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinteld, F.D., and York, R.L. (1966). *Equality of Educational Opportunity*. Washington, DC: Government Printing Office.

Dreeben, R. (1968). *On What Is Learned in School*. Reading, MA: Addison-Wesley.

Gabriel, R.M. (1985, April). *The Sustained Achievement of Compensatory Education Students: A Longitudinal Data Base*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Meyer, J.W. (1980). Levels of the Educational System and Schooling Effects, in Bidwell, C.E. and Windham, D.M., eds., *The Analysis of Educational Productivity, Volume II: Issues in Macroanalysis*. Cambridge, MA: Ballinger.

Stonehill, R.M. (1985, April). *The Sustained Achievement of Chapter 1 Students--A Summary of Findings*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.