AUTHOR Wortman, Paul M.; And Others
TITLE The First Year of the Education Voucher Demonstration: A Secondary Analysis of Student Achievement Test Scores.
INSTITUTION Northwestern Univ., Evanston, Ill. Dept. of Psychology.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE May 78
CONTRACT NIE-C-74-0115
NOTE 24p.; Research also supported by a National Science Foundation Graduate Fellowship.
AVAILABLE FROM Paul M. Wortman, Associate Director, Division of Methodology and Evaluation Research, Department of Psychology, Northwestern University, Evanston, IL 60201.
PUB TYPE Reports - Research/Technical (143) -- Journal Articles (080)
JOURNAL CIT Evaluation Quarterly; v2 n2 p193-214 May 1978

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Demonstration Programs; *Educational Finance; Educational Innovation; *Educational Vouchers; Elementary Education; *Nontraditional Education; *Program Evaluation; Quasiexperimental Design; Reading Achievement; Reading Tests; *Research Design
IDENTIFIERS Alum Rock Union School District CA

ABSTRACT

The Education Voucher Demonstration began in the Alum Rock Union Elementary School District during the 1972-73 school year. Under the voucher concept, parents freely select a school for their child and receive a credit or voucher equal to the cost of the child's education that is paid directly to the school upon enrollment. It was presumed that this form of school finance would foster competition among the schools and improve the quality of education by making schools more responsive to students' needs. An initial external evaluation at the conclusion of the first year found, however, a relative loss in reading achievement for students in the six public schools that participated in the voucher demonstration. The present report reexamines some of these data using a quasi-experimental design involving multiple pretests and individual students' test scores (rather that school means) as the unit of analysis. The results appear to indicate that the deleterious reading effect of the voucher demonstration was confined to a few within-school programs featuring nontraditional, innovative curricula. (Author)

Reprint from:

EVALUATION QUARTERLY

"The First Year of the Education Voucher Demonstration: A Secondary Analysis of Student Achievement Test Scores"

by Paul M. Wortman, Charles S. Reichardt, and Robert G. St. Pierre

SAGE PUBLICATIONS, INC
275 South Beverly Drive
Beverly Hills, California 90212

SAGE PUBLICATIONS LTD
St George's House / 44 Hatton Garden
London EC1N 8ER

The Education Voucher Demonstration began in the Alum Rock Union Elementary School District during the 1972-1973 school year. Under the voucher concept, parents freely select a school for their child and receive a credit or voucher equal to the cost of the child's education that is paid directly to the school upon enrollment. It was presumed that this form of school finance would foster competition among the schools and improve the quality of education by making schools more responsive to students' needs. An initial external evaluation at the conclusion of the first year found, however, a relative loss in reading achievement for students in the six public schools that participated in the voucher demonstration. The present report reexamines some of these data using a quasi-experimental design involving multiple pretests and individual students' test scores (rather than school means) as the unit of analysis. The results appear to indicate that the deleterious reading effect of the voucher demonstration was confined to a few within-school programs featuring nontraditional, innovative curricula.

THE FIRST YEAR OF THE EDUCATION VOUCHER DEMONSTRATION

A Secondary Analysis of Student Achievement Test Scores

PAUL M. WORTMAN
CHARLES S. REICHARDT
Northwestern University

ROBERT G. St. PIERRE
Abt Associates Inc.

In the fall of 1972, the Alum Rock Union Elementary School District in San Jose, California became the site of the first Education Voucher Demonstration. The voucher concept was designed to introduce free enterprise concepts into the educational process. This

AUTHORS' NOTE: *The work on this manuscript was supported by Contract No. C-74-0115 from the National Institute of Education and a National Science Foundation*

EVALUATION QUARTERLY, Vol. 2 No. 2, May 1978

is accomplished by allowing parents and children complete freedom in selecting their schools, by encouraging the schools to develop a variety of educational alternatives to make this choice meaningful, and, most importantly, by tying school finances to student enrollment. The program is implemented through a fiscal reorganization of the school system whereby each student receives a credit or a voucher, equivalent to the cost of his or her education, that is paid directly to the chosen school upon enrollment. With schools dependent upon these funds for their survival, they should become more responsive to student needs and, presumably, improve the quality of education these students receive. Moreover, by allocating additional compensatory funds to the vouchers of students who were eligible for the free lunch program, the schools were encouraged to be particularly receptive to the special needs of disadvantaged students.

The Alum Rock Voucher Demonstration was initially funded for a period of five years.[1] Attention, however, has been focused almost entirely on the first year. During that time only six of the district's 24 schools agreed to participate in the demonstration. In order to increase the number and breadth of choices available to parents and children, each voucher school agreed to diversify into at least two minischools which would offer varying curricular orientations, instructional strategies, and educational goals. The 22 minischools that resulted from this decision could be divided into two broad categories—those offering a traditional academic orientation ("general and specific basic academic skills") and those offering a nontraditional, innovative approach to education (i.e., "fine arts," "multi-cultural," or "activity-centered" programs).[2] According to the teachers' own estimates, while these different curricula provided a wide range of approaches to learning, there was little difference between the programs in the amount of instructional time allocated to reading and language arts (approximately 45%) in grades 3 through 6 (Weiler, 1974a: 89-90). Other differences, however, did become apparent. For example, in general, there

Graduate Fellowship. The authors thank Robert F. Boruch, Donald T. Campbell, and David A. Kenny for their helpful comments on an earlier version of this paper. We are also indebted to Superintendent William J. Jefferds and Lloyd Binen for their assistance in obtaining the data used in the analyses. Requests for reprints should be sent to Paul M. Wortman, Associate Director, Division of Methodology and Evaluation Research, Department of Psychology, Northwestern University, Evanston, Illinois 60201.

was more individualized instruction in the traditional minischool programs than in the innovative. nontraditional ones (Weiler, 1974a: 89-91).

The Rand Corporation was awarded a contract to monitor the progress of the demonstration, and in its reports on the first year (Weiler, 1976; 1974a; 1974b) examined a wide range of public policy issues including student achievement. On this latter issue, the Rand reports only examined the effects of the voucher project on reading achievement, using nonvoucher schools within the district for comparison. School means were used as the unit of analysis in statistical procedures because more disaggregated data (e.g., student scores) were thought to be unavailable.

The results from Rand's two separate analyses were contradictory. One Rand researcher (Klitgaard, 1974) mainly examined the scores from the Cooperative Primary Reading Tests (CPRT) which were administered by the state each spring. Using a gain-score analysis (see below), Klitgaard found that the performance in voucher schools dropped in the first year of the demonstration compared to both performance in prior years and performance in nonvoucher schools. The drop amounted to one-sixth of an "inter-student" standard deviation, and the same effect was found after using various measures as a control for SES differences. Another Rand analyst (Barker, 1974b) examined performance on the Metropolitan Achievement Tests, which were administered to Title I nonvoucher schools by the state and to voucher schools by Rand. Using both a correlational analysis and an analysis of gain scores, Barker found no significant differences between the performance of voucher and nonvoucher schools with and without using information on Aid to Families with Dependent Children as a control for SES differences. In interpreting these results, the Rand report (Barker, 1974a) cites a number of weaknesses in the analyses and concludes that because of these difficulties and the contradictory nature of the results, "it does not seem prudent to try to draw more than very tentative conclusions from first-year data" (p. 104).

The purpose of this paper is to reexamine these findings from the first year of the voucher demonstration. One reason for doing so is that secondary analyses, conducted from a new perspective with different assumptions, can often shed new light on the original findings and thus provide a firmer basis for decision-making. The present reanalysis, while not solving all of the problems that confronted the Rand analyses, does provide a more fine-grained look at the consequences of the voucher demonstration by focusing on the individual student's reading achievement. By examining these test scores and, in particular, those of a cohort of students who remained in Alum Rock for several years, it is possible to avoid some of the problems arising from student mobility (estimated at 30% per year). Furthermore, the use of individual scores also allows one to determine the effects of smaller and more specific components of the treatment such as the mini- (or within) school programs.

Another reason for this reanalysis was to investigate and comment on the various analytic approaches to data resulting from nonequivalent control group designs such as the voucher demonstration. This general category of designs describes a common research setting where there are pretreatment and posttreatment observations, but the individuals (or other units of analysis) under study have not been randomly assigned to treatment and control groups (Campbell and Stanley, 1966). In general, nonrandom assignment to conditions means that the treatment groups will systematically differ in both predictable and unpredictable ways. In order to reach firm conclusions concerning the effectiveness of the treatment, these selection differences must be taken into account—a situation somewhat more complicated than that encountered in the analysis of a "true" experiment.

Because the nonequivalent control group design is often employed in applied and field research settings, much has been written about the difficulties in analyzing the resulting data (Cook and Reichardt, 1976, provide an annotated guide to some of the current literature). The conclusion reached by many respected methodologists is that no completely satisfactory solution to the problems of analysis exists:

> With the data usually available for such studies, there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups [Lord, 1967: 305].
>
> If randomization is absent, it is virtually impossible in many practical circumstances to be convinced that the estimates of the effects of treatments are in fact unbiased [Cochran and Rubin, 1973: 417].

In brief, the difficulty is that in a particular research setting one specific statistical model may be appropriate for the analysis, while for other data a different statistical procedure (or a different adjustment to the same one) is required (cf. Kenny, 1975; Cronbach et al., 1977). So unless the analyst knows what model is appropriate for the specific data at hand, there is a possibility that the results from an arbitrary model will be biased. With the present lack of understanding concerning the processes that govern behavior, this knowledge will rarely be available.

Under these conditions, caution is obviously necessary in interpreting the results of such analyses. But, in spite of this caveat, it is common for prudence to yield to the pressure for hard and fast conclusions and for the shortcomings of the analysis to be ignored. For example, in subsequent reports on the voucher demonstration in more policy-relevant publications (Report on Education Research, 1974; Shanker, 1974; Warren, 1976) the "tentative conclusions" of the Rand analyses were replaced by firm statements concerning the program's failure. It is the purpose of this paper to illustrate the problems in analyzing such data and thereby to demonstrate the need for caution in reaching conclusions based on the results from the nonequivalent control group design.

METHOD

UNITS OF ANALYSIS

Students' test scores were obtained directly from the school district instead of from Rand or the State of California which had school means only. In order to track individual students (rather than schools) over time, the scores from consecutive grade levels in consecutive years were required. The only consecutive grades tested at Alum Rock were grades 1-3. Each spring (May) these grade levels were tested with the CPRT (the test on which Klitgaard's analyses found a harmful effect for the vouchers). Thus is was possible to obtain yearly test scores on a cohort of students who were in the first grade in 1970-1971, second grade in 1971-1972, and third grade during the first year of the voucher demonstration, 1972-1973. Such a sample, it should be noted, represents a trade-off between the requirements of a strong design methodology and the potential generalizability of the findings. This subsample of voucher students allows the analysis to detect the effects of smaller units than the entire school (e.g., student growth and minischool curricula), but, in doing so, the scope of the findings are restricted to the third-grade pupils (whereas Klitgaard's analysis included first, second, and sixth-graders as well).

Five elementary-grade level schools (the sixth was a middle school) joined the voucher demonstration the first year. For a nonvoucher comparison, Rand had relied primarily on the five Title I elementary schools in the Alum Rock District. For the present analysis, student data from only three of these schools were considered because the other two were racially imbalanced (as compared to the characteristics of the district) according to the California Administrative Code (Weiler, 1974a). The eight voucher and nonvoucher schools included in this sample were racially balanced. Further, voucher and nonvoucher schools appeared to be similar in other ways. The three nonvoucher schools, for example, chose to become voucher schools in the second year of the demonstration. Nonetheless, there were some obvious differences between the two groups. Specifically, Title I schools generally had more students from families on welfare than did voucher schools—51% versus 34% (Weiler, 1974a: 169).

The cohort of students who were present during grades 1 through 3 was created by matching student records across the three consecutive years. This was based on reported names, and it is likely that some spelling errors and name changes occurred in the class lists. Nevertheless, slight differences in opinion over the proper matches in a few cases did not substantially alter the results of the analyses.

Given the continuing discussion of freedom of information and rights of privacy, a brief description of the data acquisition is warranted. The recommended procedure for data release usually includes the deletion of all names and personal identifiers. In fact, legal restriction to this effect have been proposed. Such a sanction, however, would have made it impossible to construct the cohort used in the present study. Fortunately, permission for the release of the necessary information was obtained from the school district and did not require the consent of students or parents. (The names of the students, of course, were used only to create the cohort and were not used in the subsequent analyses.) Certainly the data file creation could be done by the organization that has the primary information. In this way, ethical questions concerning secondary analyses can be avoided without loss of valuable policy-relevant information.

DESIGN

With the students' actual test scores, more fine-grained comparisons were allowed in this study than in the Rand evaluation. Although the sample size of the cohort was too small to conduct a meaningful analysis of the impact of individual minischools, it was possible to examine separately the impact of the two previously described categories of minischool curricula—traditional and nontraditional—that form two nonequivalent voucher groups. In the *complete* cohort, where performance was tracked over three years, there were 354 students—150 in traditional curriculum voucher programs (from 11 mini-schools), 84 in nontraditional curriculum voucher programs (from 5 minischools), and 120 from nonvoucher programs. In analyses based on data from only the last two years, the sample size could be increased by including the scores of those students who were tested in these two years but not in the first year, 1970-1971. This adds 109 more students, for a total of 463, of which 196 were in traditional voucher programs, 103 in nontraditional, and 164 in nonvoucher. Results from this *augmented* cohort are reported only when they differ substantially from the smaller complete cohort.

Using the notation of Campbell and Stanley (1966), the basic design of this secondary analysis is described in Figure 1 where Os represent observations, Xs refer to the imposition of a treatment, and the dashed lines indicate nonrandom assignment. This design is a slight extension of the basic nonequivalent control group design since it includes an additional wave of data (grade 1 scores). Following both Director (1974) and Campbell (1974), the first two waves of data are referred to as a "dry run" quasi-experiment since the voucher program was not in effect during this time. Data from the last two years are referred to as a pretest-posttest quasi-experiment since if there is any voucher effect it might be discernible from these data.

ANALYSES

A number of statistical strategies have been suggested (and many are widely used) for analyzing data from a pretest-posttest design (Grades 2 and 3 above). One of the most common is the analysis of variance of gain scores (gain-score analysis). In this procedure, the differences between performance in grades 3 and 2 are examined to see if

[illegible]	[illegible]	[illegible]	RUN	[illegible]	
[illegible]	[illegible]	0	0		0
	[illegible]	0	0	X	0
[illegible]	[illegible]	0	0		0
		[illegible]	[illegible]		[illegible]

0 · 0' [illegible]

[illegible]

Figure 1: Design of the Educational Voucher Demonstration Secondary Analysis

one treatment group gained more than another. Any significant difference in average gain is then attributed to the effect of the treatment. Another popular strategy is the analysis of covariance (ANCOVA) where the posttest (grade 3) scores are regressed on the pretest (grade 2) scores within each treatment group separately. Significant differences between the intercepts of these regression lines are attributed to the treatment effect.

Still other analysis strategies have been devised because of the well-known fact that measurement error in the pretest biases the treatment effect estimate in the ANCOVA (cf. Campbell and Erlebacher, 1970; Cochran, 1968). Under the assumption that the ANCOVA would provide the proper results only if the pretest were measured without error, a number of corrections for the effects of using a fallible pretest in the ANCOVA have been suggested (e.g., Keesling and Wiley, 1977; Lord, 1960; Porter, 1967). The adjustment that is most convenient computationally is provided by Porter (1967). Essentially, this procedure uses an estimate of the within-group pretest reliability to regress the pretest scores toward the group means, and these adjusted pretest scores are then entered in the ANCOVA just as the unadjusted pretest scores would be.[2] Porter and Chibucos (1974) have suggested that this strategy is the most appropriate in general for data from the nonequivalent control group design. Campbell and Boruch (1975) agree with Porter and Chibucos that an adjustment to the ANCOVA is to be recommended, but disagree as to the nature of the adjustment. Under some circumstances, Campbell and Boruch argue that the proper adjustment takes account of the (within-group) pretest-posttest correlation rather than the (within-group) pretest reliability so that, computationally, one regresses the pretest scores toward the group means in proportion to this correlation rather than in proportion to the pretest reliability (also see Kenny, 1975).[3]

A more detailed discussion of these strategies is presented by Reichardt (1977), and the interested reader should also examine Cronbach et al. (1977). Suffice it to say here that: (1) each strategy imposes its own unique and specific assumptions about the state of nature, (2) under certain conditions each strategy will provide an unbiased estimate of the treatment effect, (3) under innumerable other conditions each strategy will be biased, and (4) in general one will not know which state of affairs is encountered in a specific quasi-experimental data set.

In addition to applying these strategies to the pretest-posttest data, the first wave of data (grade 1) could be used to add two other bits of information. First, each of the preceding statistical models could be applied to the dry run data. and since there was no voucher treatment at that point in time, they should, if appropriate, support the null hypothesis. Those models satisfying this criterion on the dry run data could be deemed appropriate for the analysis of the pretest-posttest data, assuming that differences between the test scores in grades 1-2 and 2-3 were due only to the introduction of the voucher programs during the latter period. Second, the data from the grades 1 and 2 can be used directly to predict the patterns of growth in the grade 3 data under the null hypothesis of no treatment effect (Boruch and Gomez, 1976). Again, it must be assumed that the same pattern of change that occurs between the first and second grades would continue on into the third grade in the absence of any voucher effect. Under this assumption, any discrepancy between the observed data in grade 3 and the prediction from grades 1 and 2 would be attributed to the effect of the voucher program. This analysis was labeled the prediction based on grades 1 and 2.

As is typical in this type of research, there was no elaborate and well-tested theory of the behavior that was under investigation nor extensive knowledge of the nature of the selection process, so it was not possible to specify which analysis strategy would be most appropriate for the cohort of students at hand. Rather than arbitrarily choosing just

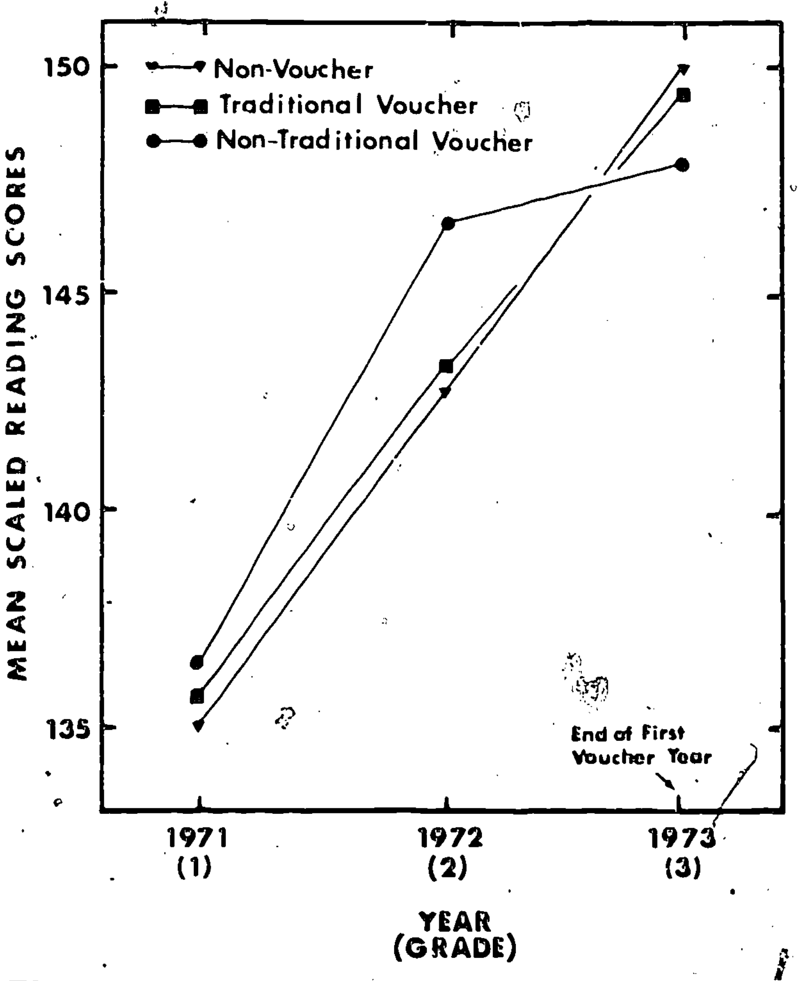

Figure 2: Mean Scaled Scores on the Cooperative Primary Reading Test for the Complete Cohort of Students Enrolled in Traditional and Nontraditional Voucher Programs and Nonvoucher Schools

one method of analysis, all of them were used. Since each statistical strategy is based on a different set of assumptions about the nature of the data, if the results of all the analyses agreed, it would indicate that the conclusions were at least robust under a range of conditions. One might also hope that this strategy of multiple tests would bracket the

TABLE 1

Summary Description of Scaled Reading Scores for the Complete Cohort of Students in the Alum Rock Union School District

Group	*Statistic*	*Grade 1* *1970-1971*	*Grade 2* *1971-1972*	*Grade 3* *1972-1973*
Traditional	Mean	135.86	143.57	149.47
N = 150	SD	4.43	9.33	9.07
Nontraditional	Mean	136.39	146.64	148.19
N = 84	SD	4.72	9.66	9.03
Nonvoucher	Mean	134.93	142.87	149.81
N = 120	SD	4.03	7.01	8.50
	$F_{(2,351)}$	2.75	5.07	<1
	p	.065	<.01	

size of the true treatment effect. In other words, while each model might be biased, the direction of the bias might be different, so that one analysis would underestimate the effect while another would overestimate it. Again, agreement among the results of these multiple analyses would lend confidence to their credibility. However, it is possible that the results of the multiple methods could all be biased in the same direction so that agreement would be misleading.

RESULTS

The first, second, and third grades had been administered a different form of the CPRT, namely, Forms 12A, 23A, and 23B, respectively. In order to make comparisons across grade levels, the raw scores were scaled according to the published norms (Educational Testing Service, 1967). The means for the three groups tracked over the three years are graphically presented in Figure 2. In Table 1 the same means appear along wih the standard deviations and F tests of the group differences at each time period. The large increase in standard deviation from the first to the second grade appears to be largely a result of the norming process. It is clear from the F tests in the table that, at least prior to the start of the voucher demonstration, the groups were (statistically) significantly nonequivalent in terms of reading ability.

A cursory look at Figure 2 reveals that there was little difference between the mean performance of the nonvoucher and the traditional voucher students over the three-year period, including the first year of the voucher demonstration. If anything, the traditional group lost ground in reading achievement compared to the nonvoucher group during the voucher demonstration's first year. In contrast, large differences are evident in comparing the mean performance of the nontraditional voucher group to the other two groups. In particular, it appears that the nontraditional group started out somewhat superior to the other two groups but lost that superiority during the first year of the demonstration. As will be seen below, the various statistical models generally agree that this latter shift should be attributed to the effect of the voucher demonstration.

PRETEST-POSTTEST

The results of the numerous tests are presented in Table 2.[4] Looking first at the pretest-posttest analyses, it is of interest to note that the results of the different models all tend to agree. Within any one row (which contains the results of a specific comparison between treatment groups), the treatment effect estimates are all in the same direction and, with a few exceptions, roughly comparable. Further, if one test reaches conventional levels of significance, the others generally do also. More specifically, the first row of the table compares the voucher programs (traditional and nontraditional combined) to the nonvoucher programs. The results reveal a significant superiority of the nonvoucher program, and are consistent with Klitgaard's original analysis based on school means which showed a decline associated with the voucher programs as a whole.

Upon disaggregating the voucher group into the traditional and nontraditional programs, a different picture emerges. There is only a small treatment effect estimate in the comparison of the traditional and nonvoucher programs (row 2) although the small difference does favor the nonvoucher group. The F values are uniformly small and nonsignificant.[5] On the other hand, it appears that the nontraditional program had a significantly harmful effect when compared to either the traditional or nonvoucher programs (rows 3 and 4). All of these analyses produce consistently large negative treatment estimates and large F ratios.

TABLE 2

Analyses of the Education Voucher Quasi-Experiments

Comparison	*Statistic*	*Gain Score Analysis*[a]	*ANCOVA*[b]	*ANCOVA with Correction for Pretest Unreliability*[b]	*ANCOVA with Pretest-Posttest Correlation Correction*[b]	*Prediction Based on Grades 1 and 2*[c]
			Pretest-Posttest			
Voucher-	β	-2.61	-1.88	-2.07	-2.64	-3.29
Nonvoucher	F	8.46**	5.38*	6.52*	10.44**	13.48**
Traditional-	β	-1.05	-.76	-.84	-1.06	-.82
Nonvoucher	F	1.14	<1	<1	1.44	<1
Traditional-	β	4.35	3.1i	3.44	4.41	6.88
Nontraditional	F	15.96**	9.92**	12.00**	19.16**	40.02**
Nontraditional-	β	-5.39	-3.88	-4.28	-5.47	-7.70
Nonvoucher	F	22.56**	14.00**	16.88**	26.54**	45.99**

Table 2 (Continued)

Comparison	*Statistic*	*Gain Score Analysis*[a]	*ANCOVA*[b]	*ANCOVA with Correction for Pretest Unreliability*[b]	*ANCOVA with Pretest-Posttest Correlation Correction*[b]	*Prediction Based on Grades 1 and 2*[a]
			Dry Run			
Voucher-	β	.68	.66	.45	-.33	
Nonvoucher	F	<1	<1	<1	<1	
Traditional-	β	-.23	-.25	-.42	-1.07	
Nonvoucher	F	<1	<1	<1	1.38	
Traditional-	β	-2.54	-2.52	-2.43	-2.06	
Nontraditional	F	6.44*	6.34*	5.88*	4.22*	
Nontraditional-	β	2.31	2.28	2.01	.99	
Nonvoucher	F	4.90*	4.68*	3.64†	<	

a. The degrees of freedom for the F tests in this column are (1,351).

b. The degrees of freedom for the F tests in these columns are (1,350).

* $p < .05$

** $p < .01$

† $p < .1$

DRY RUN

The evidence that a harmful effect is associated with the nontraditional voucher program, but not with the traditional program, is reinforced by the results in the dry run analyses. Again, the results of the models, as applied in the dry run, show general consistency within each comparison. The first two comparisons (rows 5 and 6) yield null results. Under the condition that a model must demonstrate its acceptability for a specific comparison in the dry run data (i.e., by revealing null results) before it is deemed appropriate for the pretest-posttest data, the observed null findings in the dry run strengthen the results of the corresponding comparisons in the pretest-postest. The last two comparisons in the dry run (rows 7 and 8) reveal generally large, in absolute value, treatment effect estimates (even when they appear only marginally significant), but they are in the *opposite* direction to the corresponding estimates in the pretest-postest data. This suggests, given the above assumption, that the treatment effects in the pretest-posttest data have countervailed against a natural trend in the opposite direction, and, therefore, that the results of the analyses of the former probably underestimate the size of the true effects.

Overall the results appear to be very consistent. Compared to the nonvoucher program, the traditional voucher group had no effect or at worst a slight negative effect. On the other hand, compared to either the nonvoucher or traditional voucher programs, the nontraditional voucher curricula had a reliable negative effect. The size of this estimated effect, however, is not overwhelming—roughly 5 items on a 50-item test.

More fine-grained analysis probably could provide further insight if more data were available. For example, the mean performance in both the traditional and nontraditional programs in one of the voucher schools exhibited an absolute decline from grade 2 to grade 3 and it was the only school to do so. This is somewhat surprising (assuming that the norms used to scale the test scores are reasonably appropriate) because one expects children of this age to be gradually increasing their reading ability over time. This suggests that perhaps only a few schools or minischool programs are producing the negative effects observed in the data. With this in mind, the Rand data were reexamined (Klitgaard, 1974: 108). The mean loss or decline in reaching achievement for each of the five elementary voucher schools across the first three grades was calculated by simply subtracting the mean raw-reading score for 1971-1972 from the score for 1972-1973. When this was done, this same school accounted for 42% of the total decline but only about 20% of the voucher students.

DISCUSSION

As has been noted above, conducting an appropriate analysis of the data and producing credible results is one of the major problems facing the researcher who employs a nonequivalent control group design. Two approaches have been taken to deal with this problem: (1) multiple analyses have been performed each of which have somewhat different assumptions about the nature of the data, and (2) a dry run (no treatment), double pretest has been used both directly to predict the third-grade test scores and to assess the credibility of the various statistical models. Thus for this design a strategy that employs multiple statistical models (including additional waves of data in the analysis where possible) is superior to the standard, single analysis procedure. Nevertheless, even with these safeguards, the interpretation of the results must proceed with some caution since alternative explanations for the outcomes will always be possible. Such caution usually reflects a realistic assessment of the state of the art in analyzing this quasi-experimental data and not any shortcomings on the part of the analyst.

VALIDITY OF RESULTS

For these reasons, the researcher should carefully consider the assumptions underlying the analyses and examine them in detail for plausible rival hypotheses that could produce the same results. One alternative explanation often found in educational research is that the actual pattern of growth which the individuals followed differed from the pattern that was assumed by the models. In addition to the possibility of differential growth rates, testing effects or problems of attrition also may have produced the observed results.

There is some evidence that the patterns of growth by the models were inappropriate. In particular, our use of the dry run analyses and the prediction based on grades 1 and 2 explicitly assumes similar growth patterns from grades 1 to 2 and from grades 2 to 3 (save for a treatment effect). However, children typically undergo rapid growth in academic skills during this time and it certainly would be plausible for the growth rate to vary from one grade to the next. The empirical evidence supports the contention that conditions did not remain constant over time. The within-group regression slopes of grade 2 scores on grade 1 scores were all roughly equal to 1.0, while the within-group regression slopes of grade 3 on grade 2 scores were all approximately 0.6. Moreover the distributions of test scores became more negatively skewed over time. The coefficients of skewness within the groups were approximately equal to 1.0, 0.3, and 0.4 in grades 1, 2, and 3, respectively.

Further doubt surrounds the appropriateness of the assumed growth patterns in the other analyses applied to the pretest-posttest quasi-experiment. It was hoped that the results of these multiple analyses would bracket the "true" estimate of the treatment effect. Or in other words, by using analyses with different assumptions about the pattern of outcomes, some models might underestimate the effect while others would overestimate it—forming a sort of "confidence interval." It appears plausible, however, that for these data the models are all biased in the same direction. If these models really do bracket the true effect, then within each dry run comparison the estimates should fall on both sides of the zero value since there was *no* treatment effect during this time. Unfortunately, the range of estimates encompasses the zero point in only one comparison (row 5) out of four. In addition, it is clear that the estimates in one comparison (row 7) are all substantially different from zero. Of course, this does not necessarily imply that the direction of the estimates in the pretest-posttest data are incorrect; it only suggests that they might not bracket the true value. Some of the comparisons could be underestimated, as suggested above, or some could be overestimated.

Other rival hypotheses exist as well. It is possible that a testing effect produced a spurious decline in performance in the nontraditional voucher group. The changes in the skewness of the distributions and the regression slopes of the scatterplots (noted above) indicate that a ceiling effect might have been operating at grade 3. Such an effect could artifically reduce the size of the observed differences between the groups and invalidate the results. A closer examination of the distribution of these test scores suggests, instead, that a ceiling effect probably did not account for the entire decline in the test performance of the nontraditional program relative to the other two groups. The size of the decline is simply too large to attribute it completely to an apparently slight ceiling effect.

Finally, it is necessary to consider whether differential attrition produced the observed effects. Since the performance of an intact cohort was tracked over the entire three-year period, differential attrition appears to be ruled out as a direct explanation of any change in performance. However, students learn from one another perhaps as much as from the teacher. It thus is important to determine the quality of their school environments and therefore to consider the effects of attrition among the classmates of the students in the cohort. Some trends in the data lend support to the hypothesis of differential attrition but the differences were neither large nor statistically significant.

CONCLUSIONS

What then can one reasonably conclude from this study? Despite the various threats to the credibility of the analyses and to the validity of the findings, the above interpretation of the results appears, on the whole, to be the most accurate and reasonable (although still tentative) conclusion. If the data had turned out differently, more caution would be warranted. From the present perspective, however, the trends in the data are more parsimoniously explained as a legitimate treatment effect than by a combination of rival hypotheses. Specifically, the mean performance in the nonvoucher and traditional voucher programs remain so similar over time that it is more plausible to conclude that there was only a small treatment effect (if any) than to conclude that a large effect was almost perfectly counterbalanced by some other factors. Similarly, the relative decline in performance in the nontraditional group is most plausibly attributed to a treatment effect. In educational research it is typically expected that students who are performing better than their peers will retain and even increase that superiority over time. Thus it seems more plausible to infer that the voucher demonstration was responsible for closing the gap in performance between the initially superior nontraditional voucher students and the two other comparison groups than to assume that such a pattern of growth occurred under null conditions.[6] It is unclear whether the size of this presumed effect is of educational significance.

A critical question is whether this reading loss is due to the effects of the voucher demonstration per se or to implementation of the new, innovative, nontraditional curricula. It should be remembered that the voucher concept primarily involved a fiscal and administrative reorganization in the operation of the schools and not curriculum innovation. In fact there is strong evidence (Wortman and St. Pierre, 1977) that these aspects of the voucher concept were not well implemented during the first year! Moreover, it would not be surprising if the nontraditional curricula were responsible. It is quite plausible that the loss of teacher time due to planning, developing, and modifying a new curricula produced a loss in achievement.

In light of these results it is appropriate to consider the proper role of evaluation in such an innovative program. Given the above discussion, it is clear that firm outcome assessments concerning the overall performance of the voucher demonstration are being made where none are warranted. At best this information might have been of value to administrators and teachers in developing and improving minischools. Was the voucher program entirely reponsible for the curricular choices of the bright students? Were teachers in the innovative minischools not allocating enough time to basic reading skills? These would have been relevant questions during the early stages of curiculum design and innovation that comprised a major component of the voucher demonstration. After all, the voucher program had support for at least five years. In general, the first year, or perhaps two, of a major new program should be reserved for such formative evaluation almost exclusively. It is simply too early to form any other judgment.

There is also an important issue involving the choice of instruments or observations that form the basis of the evaluation. It is likely that standard achievement tests will be employed in most educational innovations, simply because they are available and often (as in the present case) because they are mandated. The primary goal of the new voucher programs was not to increase reading or other kinds of achievement as in a compensatory program, but to increase parental choice and satisfaction with the schools. Although measures of these goals were collected and analyzed by Rand, these issues have been largely ignored in the debate over vouchers (though Levin, 1974, provides a counterexample). It is apparent that tests tailored to specific treatments and sensitive to their effects are essential to evaluation. Moreover, such measures will reduce the often encountered hostility to evaluation that results from the insensitivity of the evaluators to the goals of the program staff. Thus multiple measures, as well as multiple analyses, must be employed. Only with a diversity of dependent measures can a less distorted estimate of a program's effects be obtained.

NOTES

[1] The Education Voucher Demonstration was initially funded by the Office of Economic Opportunity and subsequently continued under the auspices of the National Institute of Education.

[2] There was no way to estimate the within-group reliability of the pretest scores from the data in our sample so we used the lower bound of the test publisher's (Educational Testing Service, 1967) alternate form estimate of reliability, which was .85. Unfortunately we can only speculate on how appropriate this value is for the sample at hand.

[3] In our sample the within-group pretest-posttest correlations were all approximately .6.

[4] Fitting curvilinear or interaction terms in the relevant analyses does not substantially alter the interpretation. Nor does the interpretation change when different units or analysis are employed. Virtually the same results were obtained in analyses based on classroom means or school means. A copy of these results is available from the author upon request. Finally, the interpretation of the results in Table 2 does not substantially change if one uses a multiple comparison technique (e.g., Scheffe, 1959) that takes into account the number of comparisons involved. The Scheffe contrasts can be generated from the tabled data by halving the F values and changing the degrees of freedom in the numerator from 1 to 2. In this way, one can insure that in examining the results from any one statistical model (but not across models), within either the dry run or pretest-posttest data, the probability of finding one *or more* statistically significant results at the .05 level is .05 (assuming the model and null hypothesis are correct).

[5] Both the ANCOVA with pretest-posttest correlation correction and the gain-score analysis reach the .05 level of statistical significance in the augmented cohort ($\beta = -1.75$, F(1,459) = 4.90 and $\beta = -1.70$, F(1,460) = 3.85, respectively).

[6] On inspecting Figure 2, one is tempted to say that the initial gap between the nontraditional voucher students and the others were *reversed* in grade 3. The results of the F tests in Table 1, however, do not support the conclusion that the nontraditional voucher program became inferior in the third grade—only that it became equivalent to the other groups.

REFERENCES

BARKER, P. (1974a) "Methodological issues in measuring student achievement," pp. 96-104 in D. Weiler (ed.) A Public School Voucher Demonstration: The First Year at Alum Rock, Technical Appendix. Santa Monica, CA: Rand Corporation, Technical Report R-1495 2-NIE (June).

——— (1974b) "Preliminary analysis of Metropolitan Achievement Test scores, voucher and Title I schools, 1972-73," pp. 120-130 in D. Weiler (ed.) A Public School Voucher Demonstration: The First Year at Alum Rock, Technical Appendix. Santa Monica, CA: Rand Corporation, Technical Report R-1495/2-NIE (June).

BORUCH, R. F. and H. GOMEZ (1976) "Double pretests for checking certain threats to the validity of some conventional evaluation designs or stalking the null hypothesis." Evanston, IL: Northwestern University, Psychology Department Technical Report.

CAMPBELL, D. T. (1974) "Measurement and experimentation in social settings." Evanston, IL: Northwestern University, Psychology Department Technical Report 627b (May).

——— and R. F. BORUCH (1975) "Making the case for randomized assignment to treatments by considering the alternatives: six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects," pp. 195-285 in C. A. Bennett and A. A. Lumsdaine (eds.) Evaluation and Experiment: Some Critical Issues in Assessing Social Programs. New York: Academic Press.

CAMPBELL, D. T. and A. ERLEBACHER (1970) "How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful," pp. 185-210 in J. Hellmuth (ed.) The Disadvantaged Child, Vol. 3. New York: Brunner Mazel.

CAMPBELL, D. T. and J. STANLEY (1966) Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

COCHRAN, W. G. (1968) "Errors of measurement in statistics." Technometrics 10: 637-666.

——— and D. B. RUBIN (1973) "Controlling bias in observational studies: a review." Sankhyā (Series A) 35: 417-446.

COOK, T. D. and C. S. REICHARDT (1976) "Guidelines—statistical analysis of nonequivalent control group designs: a guide to some current literature." Evaluation 3: 136-138.

CRONBACH, L. J., D. R. ROGOSA, R. E. FLODEN, and G. G. PRICE (1977) "Analysis of covariance in nonrandomized experiments: parameters affecting bias." Occasional Paper, Stanford Evaluation Consortium, Stanford University (August).

DIRECTOR, S. M. (1974) "Underadjustment bias in the quasi-experimental evaluation of manpower training." Ph.D. dissertation, Northwestern University.

Educational Testing Service (1967) Handbook: Cooperative Primary Tests. Princeton, NJ: Educational Testing Service.

KEESLING, J. W. and D. E. WILEY (1977) "Measurement error and the analysis of quasi-experimental data." Mehr Licht: Studies of Educative Processes, Technical Report. Chicago: CEMREL (July).

KENNY, D. A. (1975) "A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design." Psych. Bull. 82: 345-362.

KLITGAARD, R. (1974) "Preliminary analysis of achievement test scores in Alum Rock voucher and nonvoucher schools, 1972-73," pp. 105-119 in D. Weiler (ed.) A Public School Voucher Demonstration: The First Year at Alum Rock, Technical Appendix. Santa Monica, CA: Rand Corporation, Technical Report R-1495 2-NIE (June).

LEVIN, J. M. (1974) "Alum Rock, after two years: you, dear reader, have a choice." Phi Delta Kappan (November): 201-204.

LORD, F. M. (1967) "A paradox in the interpretation of group comparisons." Psych. Bull. 68: 304-305.

—— (1960) "Large-sample covariance analysis when the control variable is fallible." J. of the Amer. Statistical Association 55: 307-321.

PORTER, A. C. (1967) "The effects of using fallible variables in the analysis of covariance." Ph.D. dissertation, University of Wisconsin. (Ann Arbor, MI: University Microfilms, 1968).

—— and T. R. CHIBUCOS (1974) "Selecting analysis strategies," pp. 415-464 in G. D. Borich (ed.) Evaluating Educational Programs and Products. Englewood Cliffs, NJ: Educational Technology Publications.

REICHARDT, C. S. (1977) "The statistical analysis of data from the non-equivalent control group design," in T. D. Cook and D. T. Campbell (eds.) The Design and Conduct of Quasi-Experiments. Chicago: Rand-McNally.

Report of Education Research (1974) "What really hapened to student achievement in Alum Rock?" (September 11): 8-10.

SCHEFFE, H. (1959) The Analysis of Variance. New York: John Wiley.

SHANKER, A. (1974) "Two panaceas take it on the chin." The *New York Times* (December 1): E9.

WARREN, J. (1976) "Alum Rock Voucher Project." Educational Researcher (March): 13-15.

WEILER, D. (1976) "A public school voucher demonstration: the [illegible] year of Alum Rock summary and conclusions," pp. 279-304 in G. V. Gla[illegible] Evaluation Studies Review Annual, Vol. I. [illegible]verly Hills: Sage

—— (1974a) A Public School Voucher Demonstration: The Firs[illegible] Alum Rock. Santa Monica, CA: Rand Corporation, Technical Report R-1495-NIE (June).

—— (1974b) A Public School Voucher Demonstration: The First Year at Alum Rock, Technical Appendix. Santa Monica, CA: Rand Corporation, Technical Report, R-1945 2-NIE (June).

WORTMAN, P. M. and R. G. St. PIERRE (1977) "The Educational Voucher Demonstration: a secondary analysis." Educ. Urban Society 9 (August): 471-492.

Paul M. Wortman is Associate Director of Northwestern University's Division of Methodology and Evaluation Research, and he is a Senior Research Associate in the Department of Psychology. He and his colleagues are interested in developing new techniques of analysis, testing them by performing secondary analyses of educational and health data, and enhancing the policy relevance of evaluation methods.

Charles S. Reichardt is currently completing his graduate studies in the Department of Psychology's Division of Methodology and Evaluation Research at Northwestern University. His interests focus on both developing theoretical extensions and overcoming practical difficulties in the design and analysis of applied social research.

Robert G. St. Pierre is a Social Scientist in Education at Abt Associates Inc. of Cambridge, Massachusetts. His primary interest is in evaluation research with an emphasis on the evaluation of compensatory educational programs. He has a Ph.D. in Educational Research, Measurement and Evaluation from Boston College and spent a postdoctoral year in the Division of Methodology and Evaluation Research at Northwestern University.