DOCUMENT RESUME

ED 085 402 TM 003 343

AUTHOR	Klitgaard, Robert E; Hall, George
TITLE	Are There Unusually Effective Schools?
INSTITUTION	Rand Corp., Santa Monica, Calif.
REPORT NO	P-4995
PUB DATE	Apr 73
NOTE	37p.
AVAILABLE FROM	Rand Corporation, Publications Department, 1700 Main Street, Santa Monica, California 90406 ($2.00)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Comparative Analysis; *Educational Quality; Effective Teaching; *Evaluation; Multiple Regression Analysis; *Public Schools

ABSTRACT

A statistical analysis of data on Michigan, New York City, New York state, and Project Talent schools found evidence of schools that consistently produce outstanding students even after allowance is made for the different initial endowments of their students and for chance variation. Methodologically, like many previous studies, this report uses regression analysis of achievement data, but focuses on statistical outliers rather than central tendencies. Three tools of analysis were used to examine the residuals: (1) Histograms of residuals, showing no immediate evidence of extreme overachievers. (2) Comparisons, over different grades and years. of the number of schools that consistently over-achieved with the number expected assuming all residual variation was random. Evidence of unusually effective schools was found. (3) Comparisons of background characteristics of the hypothesized over-achieving schools with those of the average school. Outstanding Michigan schools tended to have smaller class sizes, more teachers earning over $11,000, and more teachers with greater than five year's experience. (Author)

ARE THERE UNUSUALLY EFFECTIVE SCHOOLS?

Robert E. Klitgaard and George Hall

April 1973

P-4995

The Rand Paper Series

Papers are issued by The Rand Corporation as a service to its professional staff. Their purpose is to facilitate the exchange of ideas among those who share the author's research interests; Papers are not reports prepared in fulfillment of Rand's contracts or grants. Views expressed in a Paper are the author's own, and are not necessarily shared by Rand or its research sponsors.

The Rand Corporation
Santa Monica, California 90406

ARE THERE UNUSUALLY EFFECTIVE SCHOOLS?*

Robert E. Klitgaard and George Hall

I. INTRODUCTION

Beginning with the Coleman report and continuing through the most recent research efforts,[1] scholarly analysis has eroded the belief that different school policies can lead to increases in educational achievement. Large-scale statistical studies have failed to show consistent and important relationships between what goes on in schools and variations in student learning, as measured by cognitive achievement tests.[2] To most people concerned with measuring and improving school effectiveness, these are distressing results, perhaps the most counter-intuitive findings in public policy research in the past decade.

A number of rather drastic alternatives are open. One is to accept the Coleman results and declare them the fault of the entire educational system. On this view educational effectiveness can only come about through radical reform of our whole way of schooling.

Another alternative is to reject Coleman's findings on the grounds that the wrong things were measured. One should stop reading the statisticians and economists and start reading Plato and Dewey on the true goals of education.

*The findings reported here are based on the authors' A Statistical Search for Unusually Effective Schools, R-1210-CC/RC (Santa Monica: The Rand Corporation, 1973). We are grateful to the Carnegie Corporation and Rand for research support; to Henry Acland and the University of the State of New York for data; and to Frank Berger and Gus Haggstrom for their advice and assistance. The usual caveat protecting these people and institutions from further responsibility is, of course, in order.

Or there is despair. Perhaps one should leave the educational field and go into something like bartending, where the results are clear-cut, the recipients thankful, and the emoluments more gratifying.

But there are also promising middle courses that stay in the mainstream of educational research. Without rejecting the extreme alternatives entirely, to us the most promising course seems to be in the middle; but ironically it involves getting away from central tendencies. Previous studies have indicated that *on average* school policies do not have much effect on measurable student outcomes. Suppose this is true. Might there not remain, nevertheless, a group of unusually effective schools that are different? Are there any *exceptions* to small average tendencies and insignificant regression coefficients? The mathematics of previous studies allow for such a possibility, as long as the number of exceptions is not large. In short, are there unusually effective schools?

At first glance the answer may seem obvious. Considering the enormous diversity among the nation's public schools, it would surely be incredible if some were not much better than others. Furthermore, parents and children, administrators and teachers, journalists and taxpayers seem to *act* as if some schools were unusally effective. An existence theorem seems hardly in need of proof, or even exploration.

Clearly, schools do differ in their outcomes. Some schools consistently have higher achievement scores, lower drop-out rates, more college-bound graduates, wealthier alumni, and so forth. But these results cannot be entirely attributed to the schools themselves. Pupils bring different amounts of intellectual capital to their educational experiences, in the form of different social, economic, and innate characteristics. Schools with more "advantaged" students will tend to achieve superior results. Furthermore, even when non-school background factors are identical among students in different schools, random variation will ensure that some schools will perform better than others. The question of unusually effective schools must therefore be carefully phrased: Do some schools consistently produce outstanding students even after allowance is made for the different initial endowments of their students and for chance variation?[3]

Even if unusally effective schools were rare, they would be very important for educational policy. So long as some exist and can be identified, there is hope for replication of superior performance throughout the educational system.[4] Of course, even if exemplary schools exist, it is a separate question whether their success can be reproduced elsewhere.[5] But if there are no unusually effective schools, we may have to consider seriously radically different alternatives from the present efforts of trying to discover and diffuse "best practice." We may need to make substantial changes in educational expenditures, or we may need to opt for some radical overhaul of the whole schooling system, as Silberman, Illich, and others advocate. Thus, investigating the existence of unusually effective schools is not merely a matter of scientific curiosity, but is a necessary foundation for a rational public policy towards educational improvement.

The scope of this study is limited in two ways. First, we have defined school outcomes in terms of student performance on standardized reading and mathematics achievement tests. The whole question of defining "educational effectiveness" is somehow logically prior to the search for unusually effective schools; yet we do not claim to have "solved" that problem. (It may be no more soluble than the question "what sort of house is best?") Our reliance on achievement data is not merely the result of greater availability, for we feel that such scores can reflect progress toward some valid educational objectives. But it goes without saying that test results can only be part of the story. Our paper is exploratory and conditional: if one takes achievement scores as the measure of success, is there any evidence that some schools are exceptionally successful?

The second limitation involves the questions we do not answer. There are a multitude of interesting and policy-relevant questions that can be asked about unusually effective schools. But as Sherlock Holmes properly told Henry Baskerville, the prior question is, "Does the beast exist?" The null hypothesis asserts that there are no exemplary schools. If we can discover evidence that there are, we shall leave to further researchers the detailed and important tasks of discovering why such schools exist, and how (if at all) their success can be copied.

II. PREVIOUS STUDIES

Surprisingly little research has addressed the question of unusually effective schools. Scholarly analysis has concentrated on the average effects of all school policies on educational outcomes. After controlling for student background factors, the effects of different school policies have been found to be about the same on average.[6] The anecdotal and case-study literature is replete with stories of educational successes, but the concentration is mostly on programs and not schools, is suspect of advocacy bias, and seldom includes any data.[7] The question of unusual schools has generally gone unexamined, with a few exceptions.

Part of Shaycoft's analysis of Project Talent retest data was aimed at finding out whether schools differed on their ninth-to-twelfth-grade "growth rates."[8] Not surprisingly, she found differences; but she did not control for socio-economic status (SES) or other background factors. The existence of outliers was not studied. Her study therefore did not establish that the different growth rates were due to school factors: perhaps the results were merely due to random variation and to differences in non-school variables.[9]

In their seminal work on inequality and education, Jencks and his associates provided many important analyses of school impacts.[10] Some of their findings have immediate relevance for the question of unusually effective schools--for instance, their studies of the vary narrow range of outcomes one observes among schools after controlling for various non-school factors.[11] But they did not apply the statistical tools required to determine the presence of exceptional performers.

Jencks et al.regressed school achievement scores against student background factors. The difference between the school s observed average score and the one predicted by the regression equation was the measure of whether a school was an overachiever or an underachiever. To see if there were consistent overachievers, they correlated the residuals of all schools over time. The results were unanimous: the residuals never showed a high correlation.

Correlation analysis, however, is a poor method for detecting outliers. Variations that occur throughout the entire population of schools can drown out the consistency we are interested in--that among the highest overachievers. The correlation coefficient is a measure of the strength of the linear relationship between two random variables. The relationship among the residuals (or even among the highest ones) may not be linear, yet some schools may be persistent overachievers. Even if there is no consistent tendency for all overachievers to remain that way, some may. Thus, despite the thorough and path-breaking nature of most of their work, Jencks et al do not really come to grips with our question.

An unpublished Office of Education study has come the closest to addressing our problem.[12] In 1968 Fetters, Connors, and Smith reanalyzed the Coleman data and compared the over- and underachieving schools. Figure 1 reproduces a histogram of residuals from their regression of

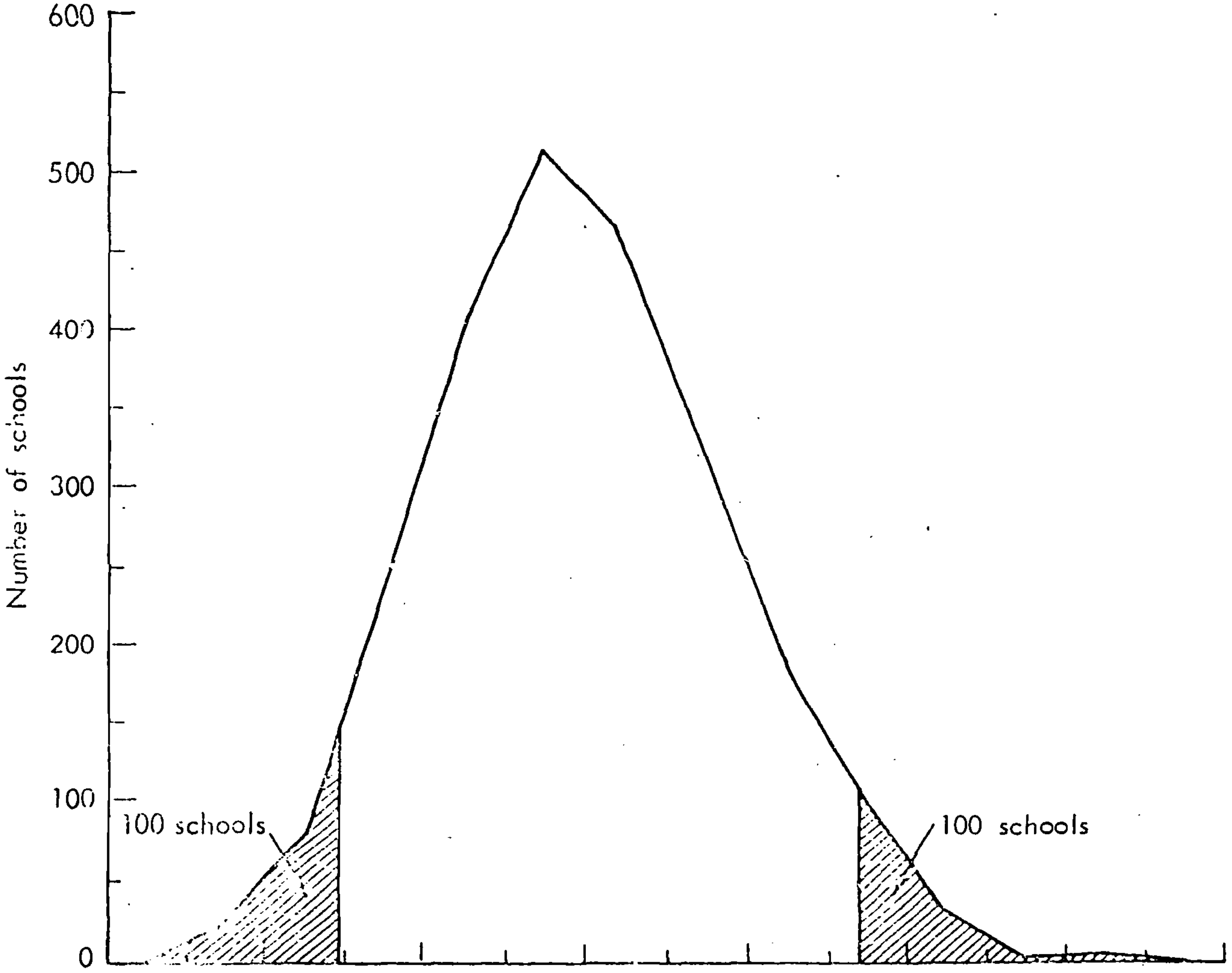


Fig.1 —Histogram of residuals from regression between SES variables and achievement scores for the EEOS data

achievement scores against various background measures for 2392 schools. Merely plotting the residuals in this fashion constitutes an important step, as one now can begin to look for evidence about the tails of the distribution and not just its central tendency. (Notice how the right tail in Figure 1 straggles: this may be a sign that there are some very exceptional performers.) But the authors went further. They compared the top 100 and bottom 100 schools, ranked by their residuals, for many input and situational characteristics. The overachieving schools tended, for example, to have more parental interest, more and better instructional equipment, smaller classes, fewer culturally and economically disadvantaged students (even after controlling for SES in the regression), less disciplinary difficulty, a better "general reputation" in the eyes of the schools' own principals, more white teachers, and a location away from industrial suburbs or the inner city.

The OE study had two important implications. First, the variables that educators had always supposed were important did distinguish between the overachieving and underachieving schools, despite the failure of these input variables to account for much variation over <u>all</u> the schools in the Coleman data. Second, the top 100 schools apparently were not on top just by chance. The fact that many school variables were significantly different between the two sets of schools is powerful evidence that the position of the top 100 schools was not a mere statistical artifact.

III. METHODOLOGY

Like many previous studies that used achievement scores as a proximate measure of school results, our basic statistical tool is regression analysis. Unlike past studies, however, we are not looking for global relationships, so we care less about characteristics of all schools and more about features of some of them. Consequently, we adopt a different approach to the regressions.

- Instead of concentrating on the properties of the regression line, the percentage of variation explained (R^2), and the coefficients of the regressor variables, we shall pay special attention to the residuals from the regression line.

- Instead of explicitly including school variables in the regression equation, we shall control only for non-school background variables and implicitly assume that what is left over after such a fit represents school effectiveness (and random variation). School effectiveness in most past studies has been measured by the size and significance of the regression coefficients of the school variables.

- Instead of including an abundance of regressor variables to explain as much variation as possible, we shall try to avoid overcontrolling.

Three reasons dictate these departures from previous practice. First, studies have shown that educational achievement is largely determined by non-school factors. This means that both school effects and purely random fluctuation have been rather small. This means that the practice of identifying school effectiveness with the residuals is not too dangerous. Residual variation could arise from a wide variety of causes besides school differences: imperfections of measurement, misspecification of the background factors, omitted variables, the choice of fitting technique, incomplete data, and the combined random fluctuations involved in all the regressor variables. But previous studies, by dint of their high R^2s, imply that such errors are not likely to be large. This does not mean, as we shall see, that we can attribute residual effects solely to schools, but from past experience we take comfort in expecting systematic errors to be small.

The second reason stems from possible intercorrelation betweeh school and background variables. If these variables suffer from multicollinearity[13] or somehow have a joint effect which cannot be attributed to school or background alone,[14] judging the true impact of s hools becomes well nigh impossible. One might reason that since we are looking for outstanding schools that are replicable, we ought to run two-stage least square regressions or specifically include an interaction term in the regression. That way, we would not call anything a "school effect" that was inextricably bound up with the background factors of the school. But this argument is inappropriate here. We do not want to prejudge the replicability question. We do not want to eliminate school effects which are intercorrelated with background effects. Furthermore, there is no convincing model of what variables should be included to capture the entire school effect. Thus, we shall use ordinary least squares and be wary of controlling for too many background factors, which might "drown out" the school effects.

The third reason we adopt our approach to regression results stems from the implications of accepting our null hypothesis. If there are no unusually effective schools, there are serious consequences for educational policy. The importance of affirming the null hypothesis means we want to be very sure that we do not accept it when it is false (we want to avoid a Type II error). If we control for a large number of background variables, there is an increased chance that through statistical interactions real outliers will not show up. Controlling for too few variables runs the risk of identifying "outliers" that could be explained by some missing regressor. However, finding no outliers under such circumstances would be a very strong result indeed. The best strategy, given the nature of our problem, is to allow exceptional schools every chance to evidence themselves by calling the entire residual the school's effect, even though this imparts an upward bias to the estimate, and by avoiding the risks of overcontrolling.

One implication of our approach is that it will be very difficult to say that outliers are the result of unusually effective schools. They may merely be the product of chance perturbations or various kinds of statistical errors. But our task may be likened to that of a detective, in contrast to the role of a judge. The detective searches for clues, the judge evaluates them. Our task is finding prima facie evidence that unusually effective schools exist, not proving their existence beyond the shadow of a doubt. If we do pinpoint some likely candidates for exceptional schools, we must realize that only after they are studied in a detailed fashion can the verdict come in.[15]

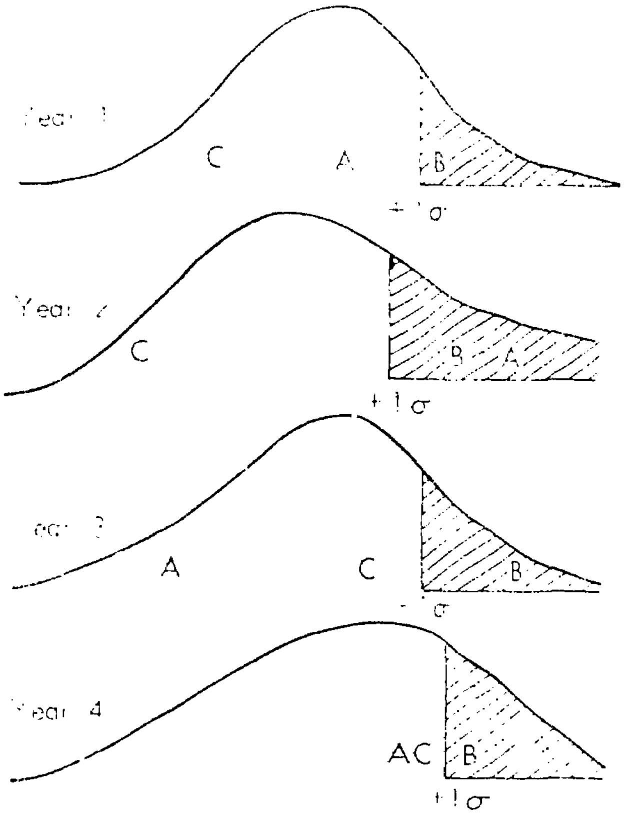
Basically, *the task is to find outliers on achievement scores that are not explained by non-school factors or random variation*. Histograms of the residuals from a regression of school achievement scores on background factors, as in Figure 1, provide a good starting point. Histograms allow easy visual inspection for "lumpiness" in the distribution of unusual tails, both of which have relevance to the question of unusually effective schools. "Lumps" would show that groups of schools are massed together in a discontinuous fashion, which may be a clue that different educational "technologies" or procedures are being used in different schools. The right tail of the histogram is of keen interest. If it is very thick, it may imply that more schools than one would expect (on the basis of a normal distribution) are performing far above average. A long tail, stretching out to four, five, and six standard deviations above the mean, is evidence that some schools are extremely high achievers. Neither "lumpiness" nor unusual right tails would constitute *conclusive* evidence of anything; but they would provide interesting clues of where to concentrate our attention.

The second tool involves looking at *series* of distributions of residuals. Each individual distribution (say, for schools in a particular year) will show the effects of random variation. A series of distributions (over many years) showing the same schools with scores consistently some distance above the mean, provides fairly strong evidence that those schools are unusual and deserve a closer look.

The null hypothesis says that all the variation in a particular distribution of residuals is a result of chance and not school effectiveness. This implies that residuals will not be correlated from year to year (as Jencks _et al_ confirmed). What we would like is some sort of "cumulative distribution" of how well schools have done over many distributions, after controlling for background factors. Then we could see if that distribution was significantly different from a theoretical distribution obtained by treating all the individual distributions of residuals as statistically independent.

We used a proxy for this cumulative distribution. All schools in a given distribution (for a particular year, say) were assigned a one if they were more than one standard deviation above the mean and a zero otherwise. Then each school's totals were added up over all the years considered, and we tested whether some schools were consistently above one standard deviation more than chance would predict.

To illustrate, assume a set of data for schools for the fourth grade during four successive years. The calculations of the proxy for the cumulative distribution are given in Figure 2, steps 1 and 2. Step 3 computes the theoretical distribution, using the binomial theorem and, in this case, a (constant) probability that a school would be more than one standard deviation above the mean in any one distribution of 0.16. Step 4 compares the actual and expected distributions using the Chi-



Procedure

1) Add totals for each school.

School	Number of Times > 1σ
A	2
B	4
C	0
.	.
.	.
.	.

2) Display frequency distribution of number of times > 1.

Number of Schools	Number of Times > 1σ
400	0
301	1
95	2
10	3
2	4

3) Using binomial theorem and assuming independence, compute theoretical frequency distribution of number of times > 1σ.

$$P(\text{x successes}) = \binom{4}{x}(.16)^{x}(.84)^{4-x}$$

4) Compare the theoretical and actual distributions, using a Chi-square test.

Number	Theoretical		Observed
0	P=.50	404	400
1	P=.38	307	301
2	P=.11	89	95
3	P=.014	11	10
4	P=.0007	1	2

Note: Chi-square 4.370 (not significant)
Degrees of freedom 4

Fig. 2 — Hypothetical illustration of analysis of sets of residuals

square test for goodness of fit. In this hypothetical case, the null hypothesis could not be rejected at the 0.05 level.

If some schools do appear to be outliers, it is important to see how they differ from the average school. Since in this paper we are only trying to discover if such schools exist and not why, the point of the comparison is not to uncover causal mechanisms, although we may find some clues. The goal is to separate random outliers from non-random ones. If many school-related characteristics of the top performers are different than the average school, it will provide strong confirmation that we have indeed located something worthy of detailed study, and not merely a statistical quirk. On the other hand, if the only differences are in non-school factors, the outliers may be the result of an omitted variable or heteroscedasticity in one of the regressors.[16]

IV. RESEARCH RESULTS

Data from three separate sources were analyzed. One was the 1969-70 and 1970-71 Michigan State school file, encompassing the fourth and seventh grades of approximately 90 percent of the state's public schools. A second involved New York City school data from 1967 to 1971, grades 2 through 6.[17] Finally, we looked at a set of 858 schools from the Project Talent high school data of 1960.

The regression equations differed from data set to data set, and we experimented with a variety of fits within the Michigan data. The Michigan equations reported here employed regressor variables of SES (derived from a student questionnaire), percent minority enrollment in the school, and community type (five categories). In the New York City data we controlled the school's mean reading score in grade k and year m for its score in grade k-1 and year m-1. Thus, for example, the fourth grade score for 1968 was regressed against the third grade score in 1967, providing a kind of measure of the students' growth from one year to the next. For Project Talent, we regressed ninth and eleventh grade composite achievement scores against an SES index. The regression results appear in Table 1.

The first surprising result was how normal-looking the individual histograms of residuals looked for all three data sources. They were all unimodally massed around the zero mean, showed no consistent or large skewness, evidenced no discontinuities, and had very well-behaved tails. The only exception was one of the Michigan series (the

Table 1

REGRESSION RESULTS

MICHIGAN SCHOOL REGRESSIONS, OMITTING RURAL SCHOOLS

Test	Equation	R^2	Standard Error	Number of Schools
R469-70	$Y = 22.18 + 4.12(\text{MIN}) + 0.50(\text{SES } 4) - 0.78(\text{C1}) + 0.27(\text{C2}) - 0.78(\text{C4})$ [illegible] [illegible] [illegible] [illegible] [illegible]	0.62	2.56	1836
M469-70	$Y = 22.32 + 4.14(\text{MIN}) + 0.50(\text{SES } 4) - 0.53(\text{C1}) + 0.02(\text{C2}) - 0.65(\text{C4})$ [illegible] [illegible] [illegible] [illegible] [illegible]	0.59	2.68	1836
R769-70	$Y = 21.00 + 2.73(\text{MIN}) + 0.54(\text{SES } 7) + 0.02(\text{C1}) + 0.46(\text{C2}) - 0.44(\text{C4})$ [illegible] [illegible] [illegible] [illegible] [illegible]	0.75	1.72	480
M769-70	$Y = 20.40 + 3.61(\text{MIN}) + 0.54(\text{SES } 7) - 0.56(\text{C1}) + 0.71(\text{C2}) - 0.44(\text{C4})$ [illegible] [illegible] [illegible] [illegible] [illegible]	0.72	2.13	480
R470-71	$Y = 22.65 + 4.13(\text{MIN}) + 0.50(\text{SES } 4) - 1.45(\text{C1}) + 0.06(\text{C2}) - 1.01(\text{C4})$ [illegible] [illegible] [illegible] [illegible] [illegible]	0.66	2.44	1891
M470-71	$Y = 20.33 + 4.39(\text{MIN}) + 0.54(\text{SES } 4) - 1.26(\text{C1}) + 0.16(\text{C2}) - 0.79(\text{C4})$ [illegible] [illegible] *33.2* *2.5* [illegible]	0.66	2.55	1891
R770-71	$Y = 20.88 + 3.89(\text{MIN}) + 0.53(\text{SES } 7) - 0.40(\text{C1}) + 0.23(\text{C2}) - 0.44(\text{C4})$ *184.5* [illegible] [illegible] [illegible] [illegible]	0.78	1.78	530
M770-71	$Y = 20.80 + 4.65(\text{MIN}) + 0.53(\text{SES } 7) - 1.84(\text{C1}) + 0.11(\text{C2}) - 0.79(\text{C4})$ *200.9* [illegible] [illegible] *0.1* *11.2*	0.78	2.04	530

NOTE: SES is based only on the school's 1970-1971 fourth- and seventh-grade scores. The minority enrollment dummy variable (MIN) has a value of 1 if percent minority > 11.3, 0 otherwise. C1, C2, and C4 are dummies for community types with those numbers. Figures in italics below the regression coefficients are the F-ratios (= t^2). R469-70 stands for the regression on reading scores for the fourth grades in 1969-1970. The other symbols are interpreted similarly.

NEW YORK CITY SCHOOL REGRESSIONS

Regression	Equation	R^2	Mean Y	Mean X	Standard Error	Number of Schools
368-267	$Y = -0.08 + 1.33X$	0.80	3.80	2.91	0.34	592
369-268	$Y = 0.54 + 1.11X$	0.77	3.70	2.84	0.34	599
370-269	$Y = 0.40 + 1.19X$	0.75	3.79	2.83	0.34	590
371-270	$Y = 0.22 + 1.15X$	0.70	3.59	2.92	0.38	591
468-367	$Y = 0.10 + 1.21X$	0.81	4.79	3.89	0.39	591
470-369	$Y = 0.71 + 1.06X$	0.78	4.66	3.72	0.39	578
568-467	$Y = 0.27 + 1.17X$	0.87	5.82	4.76	0.42	586
569-468	$Y = 0.13 + 1.14X$	0.83	5.58	4.79	0.46	592
570-469	$Y = 0.30 + 1.16X$	0.83	5.68	4.64	0.46	587
571-470	$Y = -0.15 + 1.18X$	0.79	5.38	4.67	0.51	569
668-567	$Y = 0.70 + 1.09X$	0.85	7.08	5.87	0.50	442
670-569	$Y = 0.72 + 1.05X$	0.85	6.61	5.58	0.50	444

NOTE: 368-267 refers to the regression of third-grade scores in 1968 against second-grade scores in 1967. The other symbols are interpreted similarly.

PROJECT TALENT REGRESSIONS

Test	Equation	F-ratio	R^2	Standard Error	Number of Schools	Mean Y	Standard Deviation Y	Mean SES	Standard Deviation SES
9th-grade General Aptitude	$Y = -76.37 + 5.56(\text{SES})$	307.8	0.29	66.4	746	452.7	79.0	95.2	7.7
11th-grade General Aptitude	$Y = -215.35 + 7.36(\text{SES})$	429.6	0.34	72.1	820	493.1	89.0	96.2	7.1

regressions including rural schools), which showed some slight but perhaps inconsequential thickening of the right tails. (The most deviant of these is shown in Figure 3.) We found no immediate evidence for discontinuous educational technologies nor for the existence of a few extremely high-achieving schools.

The results from looking at series of such distributions of residuals were more suggestive, although quite mixed. The Chi-square analysis results are provided in Table 2. They can be summarized as follows:

1. The Michigan data provides some evidence of unusually effective schools.

a. Counting rural schools, the Chi-square tests showed more consistently overachieving schools than chance alone would allow. For example, among the 161 schools that reported scores for all eight grade-year-test combinations, 15 were at least one standard deviation above the mean six out of eight times (less than one was expected by chance).[18] Restating these results, about 9 percent of the schools seemed able to raise their students on average by an amount equal to an increase from the 50th to the 72nd percentile, given equal background factors.[19] However, we found that most of these outstanding schools were rural and all white, even after controlling for community type and percent minority, which evidences heteroscedasticity in the control variables. By running regressions stratified on community type, we found that our regressor variables could only explain 7 percent of the

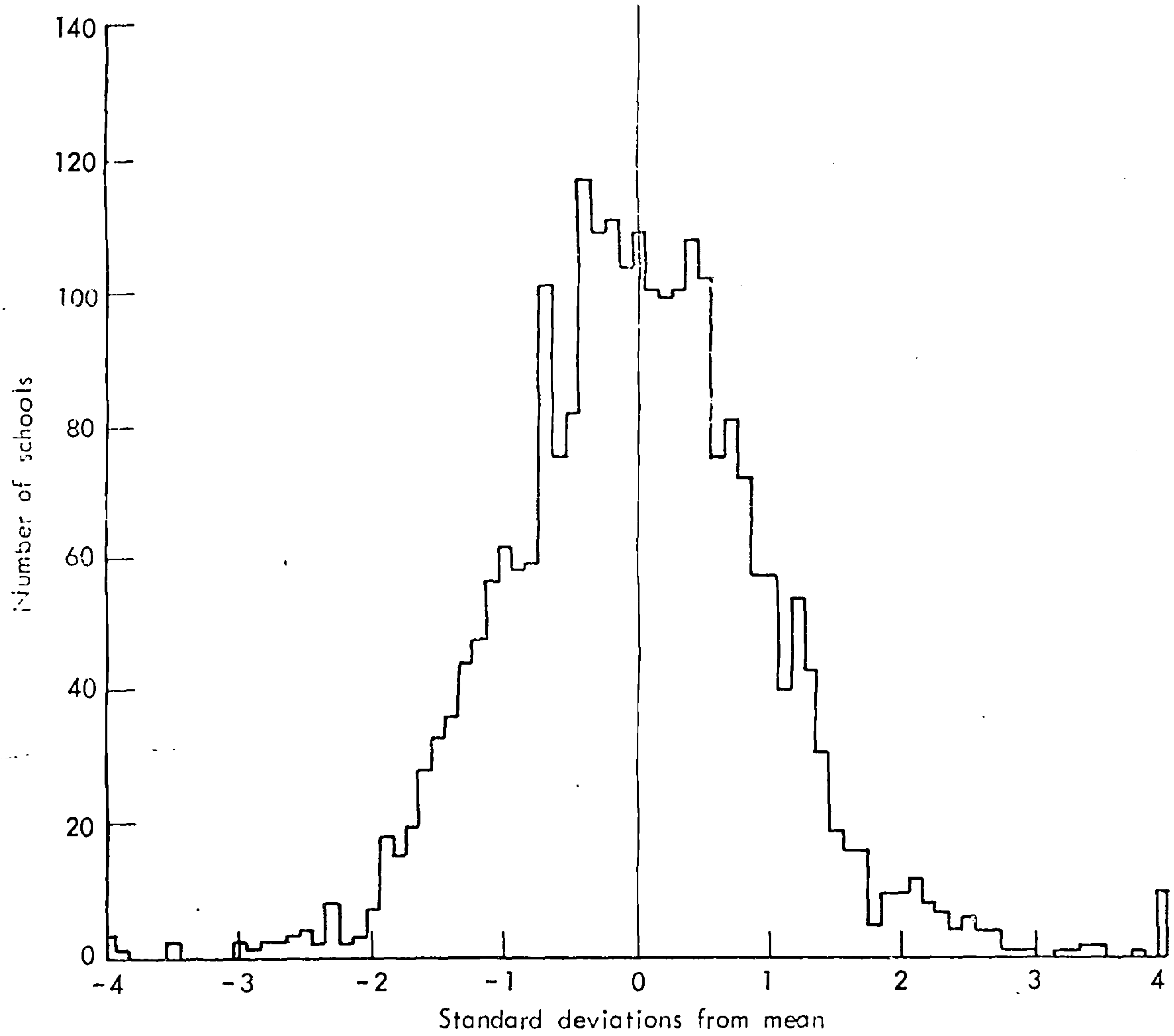


Fig. 3 —Histogram of residuals for 1969-1970 Michigan fourth-grade mathematics test, from a regression on SES, community type, and percent minority enrollment

Table 2

RESULTS OF CHI-SQUARE TESTS

MICHIGAN SCHOOLS, OMITTING RURAL SCHOOLS

Schools Reporting 8 Times			Schools Reporting 4 Times		
No. >1	Observed	Expected	No. >1	Observed	Expected
0	36	39	0	1493	1432
1	18	19	1	282	349
2	11	19	2	203	303
3	8	6	3	81	34
4	7		4	72	13
5	1				
6	1 } 14	4			
7	2				
8	3				
χ^2 = 32.6, Degrees of Freedom = 4			χ^2 = 387.2, Degrees of Freedom = 4		

NOTE: The Chi-square statistics are significant at the 0.005 level.

NEW YORK CITY ELEMENTARY SCHOOLS

Grades 3-6, 1968			Grades 3-6, 1970		
No. >1	Observed	Expected	No. >1	Observed	Expected
0	280	261	0	266	248
1	111	142	1	113	135
2	32	29	2	28	28
3 4	12	3	3 4	7	3
χ^2 = 33.6,† Degrees of Freedom = 3			χ^2 = 10.4,‡ Degrees of Freedom = 3		

Grade 5, 1967-71			Grade 3, 1967-71		
No. >1	Observed	Expected	No. >1	Observed	Expected
0	334	328	0	366	344
1	158	179	1	157	187
2	49	37	2	38	38
3 4	6	4	3 4	12	4
χ^2 = 7.8,* Degrees of Freedom = 3			χ^2 = 21.4,† Degrees of Freedom = 3		

NOTE: The probability of a school exceeding one standard deviation above the mean was approximately 0.12 for each grade/year distribution. An asterisk (*) indicates no significance at the 0.05 level. A dagger (†) indicates significance at the 0.005 level. A double dagger (‡) indicates significance at the 0.025 level.

PROJECT TALENT SCHOOLS

Grades 9 and 11, General Aptitude

No. >1	Observed	Expected
0	544	541
1	149	156
2	15	11
χ^2 = 1.5, Degrees of Freedom = 2		

NOTE: The Chi-square statistic is not significant at the 0.05 level.

variation among rural schools, compared to 50-60 percent for the other four community types combined. This may imply something about the nature of rural schools, or it may be a result of imperfect measures for SES.

b. Not including the rural schools, we also found evidence of consistent overachievers. For example, among the 2131 schools that reported scores for four grade-year-test combinations, 72 were at least one standard deviation above the mean all four times (13 were expected by chance). In other words, about 2 1/2 percent of these schools seemed able to move their students an amount equivalent to an increase from the 50th to the 65th percentile, given equal background factors.[20]

Furthermore, these 72 schools turned out to be significantly different from the average non-rural school on three out of four school-related factors. Table 3 shows that the top 72 schools tended to have smaller classes, more teachers with five or more years of experience, and more teachers earning $11,000 or more. Despite some significant differences in the number of children tested in the fourth grade, different sample sizes could not account for the position of the top 72 schools.[21] Neither could differences in non-school factors, although it was interesting to note that the overachieving schools were slightly lower than average in SES. The overachievers tended to be located more in northern Michigan than the average; once again, despite eliminating rural schools, this may be evidence for some regional/rural factor contributing to unusual effectiveness.

Table 3

COMPARISONS BETWEEN THE TOP 72 AND ALL NONRURAL SCHOOLS, MICHIGAN DATA

	Top 72 Schools[a]			All Nonrural Michigan Schools			
Variables	Mean	Std. Dev.	Number Tested	Mean	Std. Dev.	Number Tested	z-statistic[b]
School variables							
% Teachers with > 5 years' experience	67.5	16.6	72	59.1	18.6	2951	4.23†
% Teachers with master's degrees	29.2	15.1	72	28.7	15.2	2704	0.28
% Teachers earning > $11,000	51.1	21.5	72	46.2	20.4	2828	1.91*
Pupil/teacher ratio	24.5	3.5	72	25.3	3.8	2910	-1.91*
Background variables							
% Minority	12.9	21.1	72	17.0	31.3	3056	-1.61
SES 4 70	49.4	4.4	57	50.1	4.9	1923	-1.18
SES 7 70	46.9	7.8	15	49.8	4.8	555	-1.43
Number tested							
469-70	53	24	57	66	34	1986	-3.98†
769-70	208	146	17	214	133	610	-0.17
470-71	56	25	57	66	34	1984	-2.94†
770-71	237	155	15	228	137	579	0.22

NOTE: 469-70 stands for the regressions on both reading and arithmetic tests for the fourth grades in 1969-1970. The other symbols are interpreted similarly.

[a] The "top 72" schools comprise those that were at least one standard deviation above the mean four out of four times. (See Table 10.)

[b] An asterisk (*) indicates significance at the 0.05 level. A dagger (†) indicates significance at the 0.001 level. All other z-statistics are insignificant.

2. For the New York City data, the results were equivocal. We examined two years over four grades (1968 and 1970); and two grades over four years (third and fifth). Although in one year and for one grade we found some evidence of consistent overachievers, in the other year and grade it seemed that random variation could account for almost all the outliers observed. Furthermore, the consistent overachievers that were identified averaged only 1.5 inter-school standard deviations above the mean, not as large as in the Michigan schools. Very few schools indeed were above one standard deviation every time.

3. The Project Talent data showed no evidence of consistently overachieving schools apart from what chance alone would predict. This negative finding seems even stronger when one considers that only SES was used as a regressor.

In addition to looking for unusually effective schools, we took a brief look at two other levels of aggregation. Are there unusually effective districts? Using regressions by the University of the State of New York on 1969-70 and 1970-71 New York district scores for reading and mathematics,[22] we found some very suggestive evidence for outstanding districts. Among the 627 districts we studied, 30 were above one standard deviation at least five out of eight times, while less than 4 districts were expected by chance. Unfortunately, the University of the State of New York regressions did not provide information that would allow us to gauge how far these districts were able to raise their students' score in inter-student or percentile terms.

We also looked for unusually effective grades. Perhaps an entire school is not cutstanding, but certain of its grades are. However, there was little evidence to encourage further investigation of this hypothesis. The New York City results have already been discussed; there we looked at schools' third and fifth grades over time and found little evidence of consistent overachievers. No fifth grades seemed unusually effective; 2.1 percent of the third grades seemed consistently able to raise their students about half a grade level above what would have been expected given their second grade scores.

We also analyzed the Michigan data including rural schools to see if grade effects seemed greater than the school effects on both grades 4 and 7. Although there were more outstanding fourth and seventh grades than chance would predict, the amount was consistent with the notion that it was school effects rather than grade effects that accounted for these outliers.

Other levels of aggregation could of course be imagined; specifically, it would be of great interest to look for unusually effective teachers. The district findings do seem suggestive, and perhaps the search for unusual educational success should look both above and below the school level, at districts and classrooms.

V. DISCUSSION

Jencks and others have shown how tight the distribution of school achievement scores is once one controls for non-school background factors that influence such scores. Our results support that finding. We discovered no school that was consistently able to raise its students' achievement scores more than about eight-tenths of an inter-student standard deviation.[23] When we did identify a group of over-achieving schools, they comprised from 2 to 9 percent of the sample and averaged about four- to six-tenths of an inter-student standard deviation above the mean per test.[24] These schools were statistically "unusual," but whether they were unusually effective depends on one's subjective scale of magnitude. It is also important to recall that we allowed "school effectivess" to include all the variation in the residuals, not just that which could be strictly allotted to explicit school coefficients, so that our estimates of the school impacts are upwardly biased.

Nonetheless, moving away from average effects of schools does seem a worthwhile step. It appears that we have located schools deserving of further, more detailed study. It is probably also worthwhile to begin looking for unusually effective school districts and classrooms, and the methodology developed in this paper should prove useful in such efforts. As educational researchers continue to develop new measures of school outcomes, and as they begin focusing on types of students rather than school means, they should remember that most statistical techniques concentrate on the average effects of all schools. For both policy and research purposes, however, exceptions to the rule may be more important.

REFERENCES

1. Coleman, James S., *et al*, *Equality of Educational Opportunity*. Two volumes (Washington, 1966). See also Jencks, Christopher, *et al*. *Inequality* (New York, 1972) and Mosteller, Frederick and Daniel P. Moynihan, eds., *On Equality of Educational Opportunity*, (New York 1972).

2. For a review of the literature, see Averch, Harvey A., *et al*, *How Effective is Schooling?*: The Rand Corporation, R-956-PCSF/RC, Chapter III, 1972.

3. This is the so-called "valued added" approach. See, for example, Dyer, Henry S., "The Measurement of Educational Opportunity," in Mosteller, Frederick and Daniel P. Moynihan, eds., *op. cit*.

4. Ernest R. Hilgard succinctly expressed this view when he told the House Committee on Education and Labor: "We can make immediate advances in the schools by doing more widely what we already know how to do, and what more successful schools are already doing." "The Translation of Educational Research and Development into Action" in U.S. Congress, House of Representatives, Committee on Education and Labor, *Educational Research: Prospects and Priorities*, Appendix 1 to Hearings on HR 3606 and Related Bills to Create a National Institute of Education Before Select Subcommittee on Education, Washington, D. C., Government Printing Office, 1972, p. 56.

5. Reproduction might not occur for various reasons. "Production functions" may differ from school to school and area to area, hampering technology transfer. Schools' objectives may differ, especially in a decentralized educational system like the United States'. The oft-cited (if still imperfectly understood) institutional barriers to change in schools may inhibit the diffusion of superior techniques. Before these issues can be addressed, however, the prior question must be resolved: do any schools merit replication?

6. See Averch, *et al.*, *op. cit.*

7. The largest studies of exemplary educational effectiveness have been at the program level: Hawkridge, David, et al., *A Study of Selected Exemplary Programs for the Education of the Disadvantaged*, Parts I and II, American Institutes for Research (Palo Alto, 1968); Hawkridge, David, *et al.*, *A Study of Further Exemplary Programs for the Education of Disadvantaged Children*, American Institutes for Research (Palo Alto, 1969); Wargo, Michael J., *et al.*, *Further Examination of Exemplary Programs for Educating Disadvantaged Children*, American Institutes for Research (Palo Alto, 1971);and Office of the Assistant Secretary for Planning and Evaluation, *The Effectiveness of Compensatory Education*, Department of Health, Education and Welfare (Washington, D.C., 1972). George Weber looks at four "unusually effective" schools in his *Inner-City Children Can Be Taught to Read: Four Successful Schools*, Occasional Paper No. 18, Council for Basic Education (Washington, D.C., October 1971).

8. Shaycoft, Marion F., The High School Years: Growth in Cognitive Skills , Interim Report 3, American Institutes for Research Project No. 3051 (Pittsburgh, 1967), especially Chapter 7.

9. In personal communications with the authors, Ms. Shaycoft explained that she only looked "qualitatively" at the residuals, but did not plot them or correlate them. She also stated explicitly in The High School Years that she had not identified school effects: "The most that we can conclude is that entities represented by school-plus-community-plus-the-people-in-it do differ in the degree to which growth of knowledge or increase in ability occurs during the high school years, so that whether it is the school or other aspects of the neighborhood or community that bear the major part of the responsibility is a moot question." (pp. 7-11)

10. Jencks, Christopher, *et al.*, *op. cit.*

11. On the Project Talent vocabulary test, for instance, the authors calculate that the highest school increase between ninth and twelfth grades is less than 15 points, the lowest more than 5; among schools with over 20 students tested, the gains are all between 8 and 12 points. ". . . We can say that if all high schools were equally effective (or ineffective) inequality between 12th graders would fall less than one percent." (p. 90) Elementary schools appear to have a wider spread. After controlling for SES and racial composition in the Coleman data, the authors found that the top fifth of elementary schools average 10 points higher on test scores than the bottom fifth (p. 91).

12. U.S. Office of Education, National Center for Educational Statistics Division of Data Analysis and Dissemination, "Characteristics Differentiating Under- and Overachieving Schools," mimeographed (Washington, D.C., 1968).

13. See Bowles, Samuel, and Henry M. Levin, "The Determinants of Scholastic Achievement -- An Appraisal of Some Recent Evidence," Journal of Human Resources, Vol. 3, No. 1, Winter 1968, and Bowles and Levin, "More on Multicollinearity and the Effectiveness of Schools," Ibid., Vol. 3, No. 3, Summer 1968.

14. The importance of such joint or interaction effects is emphasized in Mayeske, George W., et al., A Study of Our Nation's Schools, Department of Health, Education, and Welfare (Washington, 1969).

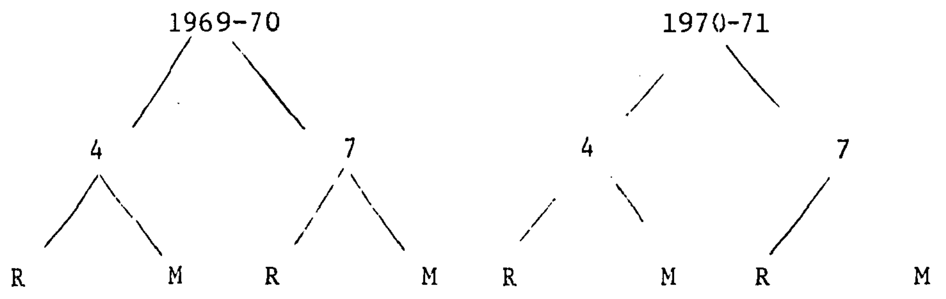
15. John Tukey used this analogy to distinguish exploratory data analysis from the more traditional confirmatory approach. Exploratory Data Analysis, Limited Preliminary Edition (Reading, Massachusetts, 1970), Vol. I, Chapter 1. Many of our methods are inspired by Tukey's philosophy.

16. Heteroscedasticity refers to non-constant variance of residuals around the regression line. If, for example, high SES schools show greater variability in their achievement scores than low and middle SES schools, most of the high (and low) outliers will be schools with high SES. The problem is distinguishing "real" socio-economic differences in residual variation from those occasioned by some property of the SES measure. This problem

arose with respect to rural schools in the Michigan data, as will be discussed below.

17. This data was graciously furnished us by Henry D. Acland of the Harvard School of Education. We would also like to acknowledge his help at a number of stages in this study. The data included about 90 percent of the city's schools, but not every school reported a score for every year.

18. A deviation from the assumption of perfect independence of the various test scores was necessary to take account of the correlation of the residuals between reading and math scores taken by the same class in the same year. The tree below shows how the eight residuals were generated:



Since the R-M residuals for a given year and grade are not independent, we reworded the null hypothesis to posit that the pairs of scores are independent.

Let X_i be the number of scores in a school's reading-mathematics pair (R_i,M_i) that exceeds one standard deviation above the mean. X_i has the possible values 0, 1, 2. Now compute a total score T_j for each school where $T_j = X_1 + X_2 \ldots . + X_j$ (j is the number of pairs of scores the school reported). Assuming the X_i are independent, one can compute null distributions for T_j using the actual probabilities of 0, 1, and 2 successes per pair. Then the actual distribution can be compared to the null distribution using a Chi-square test.

The actual probabilities for the Michigan pairs were:

		N	P(X=0)	P(X=1)	P(X=2)
Fourth-grade	69-70	1836	0.808	0.104	0.088
Seventh-grade	69-70	480	0.831	0.092	0.077
Fourth-grade	70-71	1891	0.806	0.112	0.082
Seventh-grade	70-71	530	0.832	0.083	0.085

A similar procedure was used in the New York district data.

One final note about the computation of the Chi-square statistic. In contingency tables with more than one degree of freedom, one must pool cells with small expectations in order that the Chi-square approximation be accurate. Throughout our investigations we followed a pooling rule proposed by Yarnold:

> If the number of classes s is three or more, and if r denotes the number of expectations less than five, then the minimum expectation may be as small as 5r/s.

(Yarnold, James K., "The Minimum Expectation χ^2 Goodness of Fit Tests and the Accuracy of Approximations for the Null Distribution," Journal of the American Statistical Associaton, Vol. 65, No. 330, June 1970.)

19. The 15 schools comprise about 9 percent of tne 161 that reported test scores eight times. These schools averaged two standard deviations above the inter-school mean on each test. The standard error of the regressions ranged between 2.38 and 3.94, meaning that two standard deviations was around 5-6 test points. The tests are standardized to have a mean of 50 and an inter-student standard deviation of 10; 5-6 points is therefore between five- and six-tenths of an inter-student standard deviation, which implies a change on average from the 50th percentile to about the 72nd.

20. The 72 schools averaged 1.65 inter-school standard deviations above the mean on each test, which is equivalent to about 3-5 test points, given standard errors between 1.72 and 2.68. That much of an average increase corresponds to raising an average child from the 50th percentile to about the 65th.

Are the changes documented in the last two references large? Two analogies may help. On most IQ tests, half an inter-student standard deviation is about 8 points; on the seventh grade Iowa reading test, it corresponds to almost a full grade level.

21. The number of children tested affects the estimate of the mean school score, since the standard deviation of $\bar{x} = o/\sqrt{n}$. The variation in $\bar{x}$ will be larger for smaller schools, and therefore among the outliers one would expect a more than proportionate number with small numbers of students tested. However, the statistical significance of the difference between the top 72 and average schools on number of children tests in fourth grade,

the difference between $\sqrt{53}$ and $\sqrt{66}$ is not enough to account for the magnitude of the outliers' overachievement.

22. See the University of the State of New York, New York State Performance Indicators in Education, 1972 Report (Albany, 1972), pp. 17-19.

23. The highest average over four tests was 2.92 inter-school standard deviations, corresponding to less than eight-tenths of an inter-student standard deviation.

24. Since different regressor and response variables were used, the results are not strictly comparable. However, for the same reason they may set a more convincing upper limit on the number and magnitude of unusually effective schools.