A STATISTICAL SEARCH FOR UNUSUALLY EFFECTIVE SCHOOLS

PREPARED FOR THE CARNEGIE CORPORATION OF NEW YORK

ROBERT E. KLITGAARD
GEORGE R. HALL

R-1210-CC/RC
MARCH 1973

Rand
SANTA MONICA, CA. 90406

A STATISTICAL SEARCH FOR UNUSUALLY EFFECTIVE SCHOOLS

PREPARED FOR THE CARNEGIE CORPORATION OF NEW YORK

ROBERT E. KLITGAARD
GEORGE R. HALL

R-1210-CC/RC
MARCH 1973

Rand
SANTA MONICA, CA. 90406

PREFACE

Much recent research has cast doubt on the effectiveness of public schools. Investigators have failed to identify educational policies that consistently and importantly affect measurable scholastic or occupational performance--certainly one of the most surprising findings in any public policy area in the last decade. But most research has suffered from two shortcomings: an almost exclusive reliance on achievement scores as the measure of school effectiveness, and a concentration on the *average* effects of policies over *all* schools. This report, funded by Rand and the Carnegie Corporation of New York, addresses the second shortcoming.

Given a particular definition of school effectiveness, where should one look for educational success? Our guiding principle is to search for effectiveness in the exceptional school instead of the average.

This study comprises an exploratory analysis of six large sources of data on public schools. Different grades, regions, units of analysis, control variables, and years are used: the goal is not uniformity of analysis but the discovery of evidence that some schools are unusually effective. To locate such schools, we employ statistical methods applicable in other research that wishes to identify exceptional performers.

This study has three important implications for research and public policy. Methodologically, it emphasizes getting away from central tendencies when searching for educational success. Statistically, it discusses the shortcomings of previous methods of analysis and provides appropriate exploratory techniques for the identification of outstanding schools. Empirically, it locates possibly overachieving schools and districts that deserve further research to discover why they are exceptional and how their success might be replicated elsewhere.

SUMMARY

The existence of unusually effective schools is an important question for educational policy. If exceptional schools do exist, they may be replicable elsewhere; if they do not, we may need to consider radical changes in the educational system. This report concentrates on the following question: Do some schools consistently produce outstanding students even after allowance is made for the different initial endowments of their students and for chance variation?

Surprisingly, this issue has largely been ignored by researchers. They have tended to look for educational success in the average effects of policies over all schools, instead of the exceptional effects of policies in some schools.

Methodologically, like many previous studies, this study utilized regression analysis of achievement data, but with a completely different emphasis. Since we were concerned with outliers rather than central tendencies, we focused not on the regression line and its coefficients, but on the residuals from that line. Since there is no accepted model of the school variables that affect scholastic or educational performance, and since previous studies have shown that most interschool variation is explained by variation in nonschool factors, we controlled only for nonschool factors like socioeconomic status, minority enrollment, regional variables, and previous test scores. Then we assumed that all residual variation represented the school effects (and random fluctuation). Since our task was exploratory, we experimented freely with simple and complicated controls, using different data and different kinds of fits. We did not want to miss exceptional schools, if they existed. Therefore, if we located unusual schools, we could not definitely say whether their performances were due to school policies or not; but if we found no outliers, we would have a very strong result indeed. In effect, we defined upper limits on the probable number and magnitude of exceptional schools.

Three tools of analysis were used to examine the residuals. First, we examined histograms of residuals to discover the presence of extremely overachieving schools and different educational "technologies" among

groups of schools. Second, using scores from different grades and years, we compared the number of schools that consistently overachieved with the number expected assuming all residual variation was random. Third, when possible, we contrasted policy-related and background characteristics of the overachieving schools with those of the average school. Differences in policy-related factors would indicate that outliers were indeed not random; but differences in background variables could indicate statistical problems like heteroskedasticity, curvilinearity, or an omitted control.

We examined six large data sets, two of which proved to be unhelpful. Our results can be summarized as follows:

1. The distributions of school achievement scores were always extremely tight, once nonschool background factors were held constant. The histograms of the residuals were uniformly normal-looking, with well-behaved tails and no discontinuities. Thus, there was no immediate evidence of extreme overachievers or "gaps" in educational technology.
2. We located evidence of unusually effective schools, however, especially in the 1969-1971 Michigan data. When such schools were identified, they composed from 2 to 9 percent of the sample and averaged about 0.4 to 0.6 interstudent standard deviations above the mean, corresponding to an increase from the fiftieth to the sixty-fifth or seventieth percentile. (Because different response and regressor variables were used, the different results are not strictly comparable.) Using New York data, we also found evidence of unusually effective *districts*. In other analyses, notably with Project Talent data, we discovered no outstanding schools apart from those chance alone would allow. No matter how simple the controls and even assuming all residual variation reflected school effectiveness, no school in any data set was consistently able to raise its students' achievement scores more than about 0.8 interstudent standard deviations.
3. Outstanding Michigan schools tended to be rural and white,

even after controlling for region and racial composition. Heteroskedasticity was found. After eliminating rural schools from consideration and using a minority enrollment dummy, the top Michigan schools were found to have smaller class sizes, more teachers with five or more years of experience, and more teachers earning $11,000 or more annually. These results provided strong evidence that such schools were not statistical quirks.

The policy implications of this report depend on the appropriate scale of magnitude. Is half an interstudent standard deviation "a lot"? We discovered schools that were statistically "unusual," but whether they were unusually *effective* is a question transcending mathematics. Nonetheless, consistently better schools were located that probably deserve in-depth study. A methodological lesson is also important: we ought to be getting away from average effects. For both policy and research purposes, exceptions to the rule may be more important.

ACKNOWLEDGMENTS

We are grateful to the many people who provided assistance and advice during the course of this study. H. Acland of the Harvard Graduate School of Education supplied the data on New York City elementary schools as well as guidance at many stages of the project. D. Hoaglin of the Harvard Statistics Department and T. Shemo of the Harvard Public Policy Program gave us early statistical and computational assistance. M. Shaycoft of the American Institutes for Research discussed the results of her work on the Project Talent data with us. J. Maybee, A. Stewart, and G. Wohlferd of the University of the State of New York (the Department of Education) furnished us with two sets of New York state data.

We are also indebted to many members of the Rand staff. F. Berger performed most of the computer applications. We also thank G. Bass, S. Carroll, G. Carter, L. Colbert, G. Haggstrom, C. Johnston, J. Pincus, W. Rogers, J. Rolph, K. Scofield, and D. Weiler for their help.

The usual caveat protecting these courteous people from further responsibility is, of course, in order.

CONTENTS

PREFACE iii

SUMMARY v

ACKNOWLEDGMENTS ix

FIGURES xiii

TABLES xv

Section
I. INTRODUCTION 1
The Search for Educational Effectiveness 1
Achievement Scores and School Effectiveness 3

II. MAJOR PREVIOUS STUDIES 6

III. STATISTICAL ANALYSIS 16
Methodology 16
Michigan School Data 22
Preliminary Explorations 23
Controlling for Additional Variables 32
Heteroskedasticity Problems 43
Characteristics of Exceptional Schools 49
New York City Data 53
Project Talent Data 57
New York State and Yardstick Project Data 59
Alternative Levels of Aggregation 60

IV. CONCLUSIONS 67

BIBLIOGRAPHY 73

FIGURES

1. Three Different Hypothetical Scatter Plots with 0.65 Correlation Coefficients 10

2. Part of a Hypothetical Distribution of Residuals 11

3. Part of a Hypothetical Distribution of Residuals Showing Persistence of Outliers 12

4. Histogram of Residuals from Regression Between SES Variables and Achievement Scores for the EEOS Data 15

5. Example of a Histogram of Residuals 19

6. Hypothetical Illustration of Analysis of Sets of Residuals 21

7. Histogram of Residuals for 1970-1971 Michigan Fourth-Grade Mathematics Test, After Controlling for Socioeconomic Status 25

8. Histogram of Residuals for 1970-1971 Michigan Fourth-Grade Reading Test, After Controlling for Socioeconomic Status 26

9. Histogram of Residuals for 1969-1970 Michigan Fourth-Grade Mathematics Test, from a Regression Controlling for Racial Composition, Community Type, and Socioeconomic Status 35

10. Histogram of Residuals for 1970-1971 Michigan Seventh-Grade Mathematics Test, from a Regression Controlling for Racial Composition, Community Type, and Socioeconomic Status 36

11. Plot of Achievement Scores and Number of Students Tested, Random Sample of 400 Michigan Schools (R470-71) 42

12. Plot of Achievement Scores and Percent Minority, Random Sample of 400 Michigan Schools (R470-71) 44

13. Schematic Representation of Interschool and Interstudent Variation of Achievement Scores 68

TABLES

1. Regressions on Michigan Data, 1969-1970 and 1970-1971, Fourth and Seventh Grades, Reading and Mathematics Tests 23

2. Results of Chi-Square Tests of Differences Between Observed and Expected Distributions of Residuals for Michigan Schools 30

3. Comparisons Among the Top 100, Middle 100, and Bottom 100 Michigan Schools Ranked by Residuals on the Fourth-Grade Reading Test, 1970-1971 31

4. Michigan School Regressions, Controlling for Racial Composition, Community Type, and Socioeconomic Status 34

5. Number of Expected and Observed Michigan Schools with Large Residual Test Scores after Controlling for Racial Composition, Community Type, and Socioeconomic Status 37

6. Results of Chi-Square Tests of Differences Between Observed and Expected Distributions of Residuals for Michigan Schools, Revised Regressions 39

7. Characteristics of 15 Overachieving Michigan Schools ... 41

8. Michigan School Regressions Stratified by Community Type, Fourth-Grade Reading Test, 1970-1971 45

9. Michigan School Regressions, Omitting Rural Schools 46

10. Results of Chi-Square Tests of Differences Between Observed and Expected Distributions of Residuals for Michigan Schools, Omitting Rural Schools 48

11. Comparisons Between Top 72 and All Nonrural Schools, Michigan Data 51

12. Geographical Comparisons Between Top 72 and All Nonrural Schools, Michigan Data 52

13. New York City School Regressions 54

14. Results of Chi-Square Tests of Differences Between Observed and Expected Distributions of Residuals for New York City Elementary Schools 56

15. Project Talent School Regressions 58

16. Results of Chi-Square Test of Difference Between Observed and Expected Distributions of Residuals for Project Talent Schools 58

17. New York State School District Regressions 61

18. Results of Chi-Square Test of Difference Between Observed and Expected Distributions of Residuals for New York State School Districts 63

19. Comparing Grade Effects and School Effects for the Michigan Data 64

I. INTRODUCTION

THE SEARCH FOR EDUCATIONAL EFFECTIVENESS

Beginning with the path-breaking Coleman report (1966) and continuing through the most recent research efforts, scholarly analysis has eroded the belief that different school policies can increase 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 (Averch, et al. 1972). To most people concerned with education, these are surprising and distressing results. They countermand common sense and seem to imply that belief in effective education is a delusion.

In the wake of such research, where should one look for educational effectiveness? Many educational researchers have accepted the Coleman results and concluded that effective schools do not exist, though they draw quite varied policy lessons. Some say that because no educational policy is particularly successful, less money should be spent on schools; others argue that more should be spent, in order to make up for low productivity; and still others call for a radical overhaul of the entire ineffective system. Many educators, however, have rejected the Coleman findings on the basis that the wrong outcomes are being measured. Educational effectiveness has not been found, they say, because it has not been correctly measured; achievement scores are imperfect statistics relating to only a small facet of what education is about.

However, perhaps educational research has looked in the wrong places for evidence of effectiveness. Previous studies have indicated that, *on average*, school policies do not greatly affect measurable student scholastic and occupational performance. Suppose this is true. Might there not remain, nevertheless, a group of 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?

Considering the enormous diversity among the nation's public schools, it initially seems obvious that some should be much better than others. Furthermore, parents and children, administrators and teachers, journalists and taxpayers seem to *act* as if some schools were unusually effective. And, clearly, schools do differ in their performances. Some consistently have higher achievement scores, lower drop-out rates, more college-bound graduates, wealthier alumni, and so forth.

These differences, however, cannot be attributed entirely to the schools themselves. Pupils bring different amounts of intellectual capital to their educational experiences, due to varying social, economic, and personal characteristics. Schools with privileged students will tend to achieve superior results. Furthermore, even when socioeconomic and personal 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?*[*]

This is an important policy question. Even if unusually effective schools are rare, so long as some exist and can be identified there is hope that their superior performance might be replicated throughout the educational system.[†] But if no exceptional schools exist, we may have to consider alternatives radically different from present attempts to discover and diffuse "best practice." We may need to make substantial changes in educational expenditures, or even overhaul the entire educational system. Thus, investigating the existence of unusually effective

---

[*]Note that unusual effectiveness is a relative concept here. An unusually effective school achieves educational results for its students that are superior to the average results achieved by schools with student bodies with similar characteristics. This concept, sometimes called the "value-added" or "gain" approach to school effectiveness, is set out in a paper by Dyer in Mosteller and Moynihan (1972). See also Barro (1970).

[†]Hilgard (1972) succinctly expressed this view when he told the House Committee on Education and Labor, "We can make immediate advances

schools is not merely a matter of scientific curiosity, but a necessary foundation for a rational public policy toward educational improvement.

The scope of our empirical investigation of this question is limited in two ways. First, in this exercise we used only achievement scores to measure effectiveness. After initial experimentation, we chose not to use other measures of student achievement or alternative methodological approaches, because of their insufficient conceptual development and lack of appropriate data. Our studies are thus exploratory and conditional: If one takes as the measure of success achievement scores controlled for nonschool background factors, does one find any evidence that some schools are exceptionally successful?

Second, we do not answer a multitude of interesting and policy-relevant questions that can be asked about unusually effective schools. As Sherlock Holmes properly reminded 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 future researchers the detailed and important work of discovering why they exist and how (if at all) their success can be copied.

ACHIEVEMENT SCORES AND SCHOOL EFFECTIVENESS

Using achievement scores to measure educational effectiveness is one of a number of approaches to evaluation that regard schooling as a means to some end or set of ends. The "ends" may be viewed as proximate or ultimate, one-time or continuous, quantitative or qualitative, economic or academic; but they share the property of being *products* or *outcomes* of education.

---

in the schools by doing more widely what we already know how to do, and what more successful schools are already doing" (p. 56).

On the negative side, replication might not occur for several reasons. Educational "production functions" may differ from school to school and area to area, hampering technology transfer. Educational objectives may also vary, especially in a system emphasizing local control. Also, the oft-cited (if still imperfectly understood) institutional barriers to change in schools may hamper the diffusion of superior techniques. Our empirical efforts ignore these questions and ask, "Do any schools exist that because of unusual success merit replication?"

This "educational-outcome" approach has stimulated investigators for many years, but it has also drawn its share of censure. Many critics have protested an implicit factory model, as if to look for outcomes were to treat students like raw materials, teachers like machine operators, and knowledge like a production good. Many such general criticisms miss the mark. An outcome *approach* can be adapted to any process or phenomenon. It is a purely logical scheme by which the process is structured for certain kinds of analyses; the structure itself is quite innocuous. The real issue is the choice of products to be studied and the values assigned to them. Most educators who disagree with the educational-outcome approach do so because they feel the wrong results are given preference, or some important aspect of the process itself is not being counted as an outcome.

Different kinds of educational products have been analyzed, including later educational success, future income, and future socioeconomic status, as well as more proximate measures like drop-out rates, grades, and achievement scores. Long-term studies take so much time to complete, and the problem of linking an individual's history to his education is so difficult, that most researchers have relied on proximate measures.* Of these, the most popular have been achievement scores.

The use of achievement tests has always been controversial; when they became the chief indicators of educational effectiveness, it was no surprise that considerable criticism arose. Some testers themselves, perhaps embarrassed by the sudden seriousness with which their results were being taken by policymakers, called IQ and grade equivalency scores "monstrosities."† Government data collections and large-scale surveys--so long sought by educational researchers and designed by their peers and leaders--were now decried as incomplete, biased, plagued by attrition and turnover: in short, as no basis whatever for policy decisions.‡

---

*For recent examples of the use of ultimate measures, see Jencks, et al. (1972), Bowles and Gintis (forthcoming), and the ongoing work of the American Institutes of Research on Project Talent data.

†See, for example, the remarks of Henry S. Dyer, vice president of the Educational Testing Service, at the conference entitled "Educational Accountability" held in Washington, D.C., March 23, 1971 (quoted in Stevens 1971).

‡See, among many such, Bowles and Levin (1968) and Guthrie (1972).

And many critics said that achievement scores were simply the wrong measure of the educational product. They indicated only a small part of what schools were about, and that part incompletely.

Unfortunately, attractive alternatives are lacking. After preliminary efforts with other outcome measures, we have grudgingly confined ourselves to achievement scores, though our use of such information was not merely the result of its greater availability. We feel that test results can reflect progress toward valid educational objectives. Our investigations can thus be viewed as an assessment of the value of using achievement data to locate unusually effective schools. In this research, we examined six data sets:

1. Michigan schools, 1969-1970 and 1970-1971, grades 4 and 7.
2. New York City elementary schools, 1967-1971, grades 2 through 6.
3. Project Talent data, 1960, grades 9 and 12.
4. New York state school districts, 1969-1970 and 1970-1971, grades 3 and 6.
5. New York state schools, 1966, grade 1; 1968, grade 3; and 1971, grade 6.
6. Project Yardstick data, many years and grades.

The rest of the report is divided into three parts. Section II describes major previous studies of educational effectiveness. Section III outlines the methodology we used to locate outliers,* and reviews and assesses the relevance of previous studies of achievement scores to the search for unusually effective schools. Section IV presents the results of our data analyses and discusses their implications.

---

*"Outliers" are unusual data points. John Tukey has divided "outliers" into two categories--"outside values" and "detached observations"--and has given them more or less quantitative definitions (1970, Ch. 5, p. 17). In this report, the term will be used qualitatively to indicate observations that are exceptional compared with the data's central tendency and overall distribution.

II. MAJOR PREVIOUS STUDIES

Large regression studies have concentrated on the *average* effects of school variables on achievement. These effects have generally been found to be small. Coleman, et al. (1966)--and later Jencks and Smith in papers in the Mosteller and Moynihan (1972) reanalysis of the data from the Equal Educational Opportunity Survey (EEOS)--stratified schools by region, race, and degree of urbanization, but their emphasis was still on average effects within these categories. The possibility of unusual performers at the top of the distribution has received surprisingly little attention.

Shaycoft's (1967) analysis of Project Talent data is a partial exception. She examined data on a 1963 retest of a cohort of some 7500 high school seniors who had previously been examined as ninth-graders in the Project Talent survey. She wanted to see if schools' "growth rates" on achievement tests during this period were the same. Not surprisingly, she found that schools differed. However, socioeconomic status (SES) and the other background factors were not controlled, so one cannot infer that the observed spread was due to school factors alone.* Further, she did not correlate the growth rates on different tests across schools, so it was impossible to ascertain whether any schools were uniformly excellent. Finally, she concluded that

> there was no clear relation to be seen between test performance and such school characteristics as course offerings. Schools that reported a strong program in mathematics or in physical science (many courses offered, and strong requirements for graduation) were not especially likely to be among those that scored higher than average in those areas. (Pp. 7-24f.)

---

*Shaycoft recognizes this shortcoming explicitly: "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" (p. 7-11).

Discriminant functions used to see how "schools together with other environmental factors acting on the set of students considered as a whole differ in their effects during the last three high school years" (p. 7-23) revealed a wide array of dimensions along which schools and communities differed. However, the question of outliers was not addressed. In personal communications with the authors, Shaycoft explained that she looked "qualitatively" at the residuals, but did not actually plot them or analyze them statistically.

In sum, Shaycoft's study showed that schools do differ significantly in their growth rates on achievement tests. However, she did not establish that the differences were due to school factors; the results she obtained may have been due to differences in socioeconomic variables.

In their seminal work on inequality and education, Jencks and his associates (1972) performed many important analyses of educational impacts. Some of their findings have immediate bearing on the question of unusually good schools--for instance, their studies of the range of outcomes one observes after controlling for different socioeconomic and personal factors. But they did not apply the proper statistical tools to determine the presence of exceptional schools.

Jencks and his associates found that the *range in outcomes*, once controlled for socioeconomic status and other background variables, *is small*.

> We did not expect all schools to have precisely the same effects on children's test scores. The differences are, however, surprisingly small. This suggests that public schools are more alike than parents and teachers think they are. (P. 89.)

On the Project Talent vocabulary test, for instance, they calculated that the highest school has less than a 15-point increase between the ninth and twelfth grades and the lowest more than 5. Among schools with over 20 students tested, the gains are all between 8 and 12 points (p. 90). Furthermore,

> if we average schools' effects on several different tests, the average twelfth grader's overall performance is within 3 points of what we would expect on the basis of his ninth grade scores. Stating it slightly differently, we can say that if all high schools were equally effective (or ineffective) inequality between twelfth graders would fall less than one percent. (P. 90.)

Similar results were derived from suitably controlled data compiled by the Equal Educational Opportunity Survey. The authors found that the average difference between a high school's mean ninth-grade scores and its mean twelfth-grade scores was 3 points (p. 175), although they did not cite evidence on the range of these differences. Elementary schools appeared to have a larger spread. After controlling for socioeconomic status and racial composition in the EEOS data, the authors found that the top fifth of the elementary schools averaged 10 points higher on test scores than the bottom fifth (p. 91).

The research by Jencks and his associates shows that the distribution of school achievement scores is tight after one controls for the socioeconomic factors that influence such scores. If outliers exist, they will not be large when compared with intraschool variation.

To determine whether there were consistent overachievers, Jencks and his associates examined the residuals from regressions against SES and student background variables to see if the residuals correlated over time and over different kinds of tests. Their results were unanimous and discouraging: the residuals never showed a high correlation. Jencks and his associates judged that the low correlation ruled out the existence of consistently exceptional schools.

Correlation analysis, however, is a poor tool for detecting outliers. Variations that occur throughout the entire population can drown out the variation among exceptional schools. The correlation coefficient is a measure of the strength of the linear relationship between two random variables. The relationship among the residuals may not be linear, yet there may still be evidence of consistent overachievers. Moreover, even if all overachievers do not maintain their excellence, some may.

The following examples clarify why correlation is not an appropriate tool for locating exceptional schools.

1. *Variation throughout the entire population enters into the calculation of a correlative coefficient.* Thus, three seemingly very different distributions can have the same correlation coefficient. Figure 1 shows three different scatter plots with the same correlation coefficient between the X and Y variables, but with quite different appearances.
2. *Correlation measures the linear relationship.* A school's residuals could remain "far out" (say, past two standard deviations), but retain substantial variability beyond that point. The correlation coefficient between its successive scores might be low. Yet if we defined "unusually effective" as overachieving more than two standard deviations beyond that predicted by student background variables, the school would be consistently exceptional. That is, a school could be repeatedly outstanding and still have a nonlinear relationship among its residuals. This extremely important point is illustrated in Fig. 2, which gives a two-year plot of part of a hypothetical set of residuals from a regression of school achievement scores. The correlation of year-to-year residuals is low. However, certain schools are exceptional in both years. Despite the nonlinear relationship among the residuals as a whole and among the outliers in particular, visual inspection shows clearly that some schools are persistent overachievers.
3. *There may be little tendency, on average, for high-scoring schools to remain exceptional, yet some schools may be overachievers time after time.* Even a statistic based on all schools that exceed some standard deviation defining excellence, therefore, may still run aground. Figure 3 displays part of a hypothetical distribution of schools scoring two or more standard deviations above the mean at least once. The correlation among these schools' residuals is not high. Most schools are outliers for only one year, but some remain outside in both years.

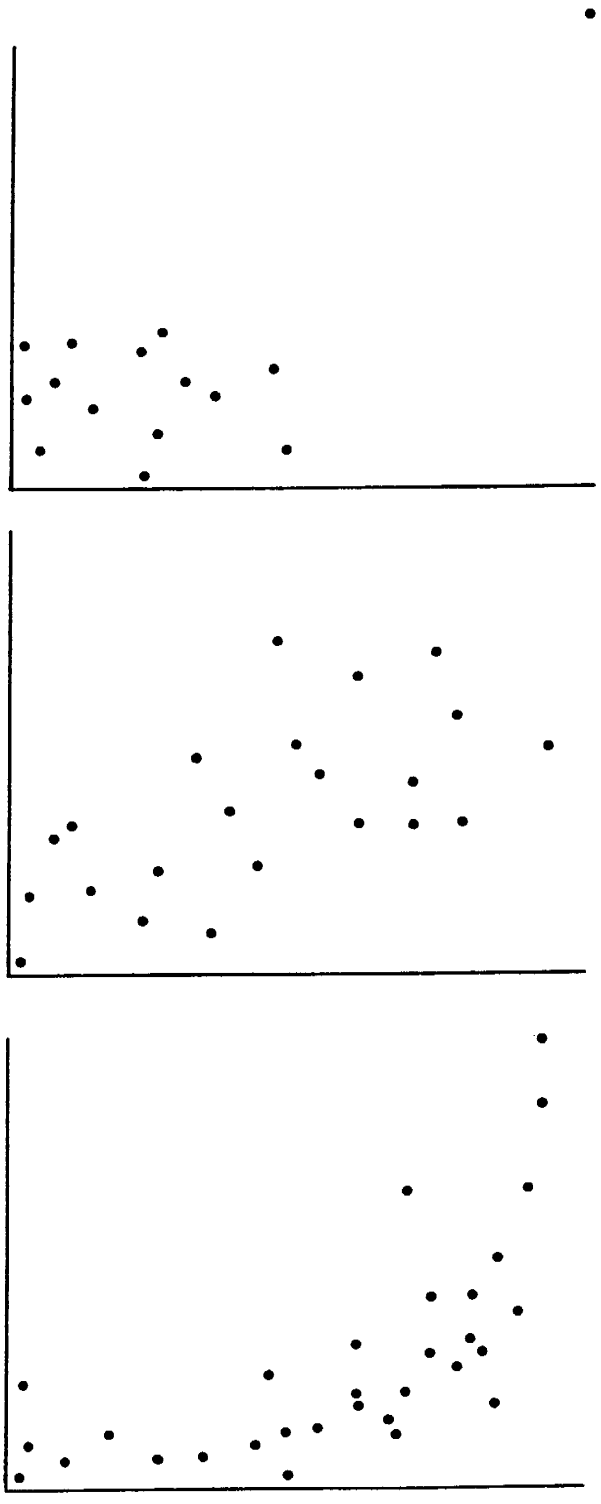


Fig. 1 — Three different hypothetical scatter plots with 0.65 correlation coefficients

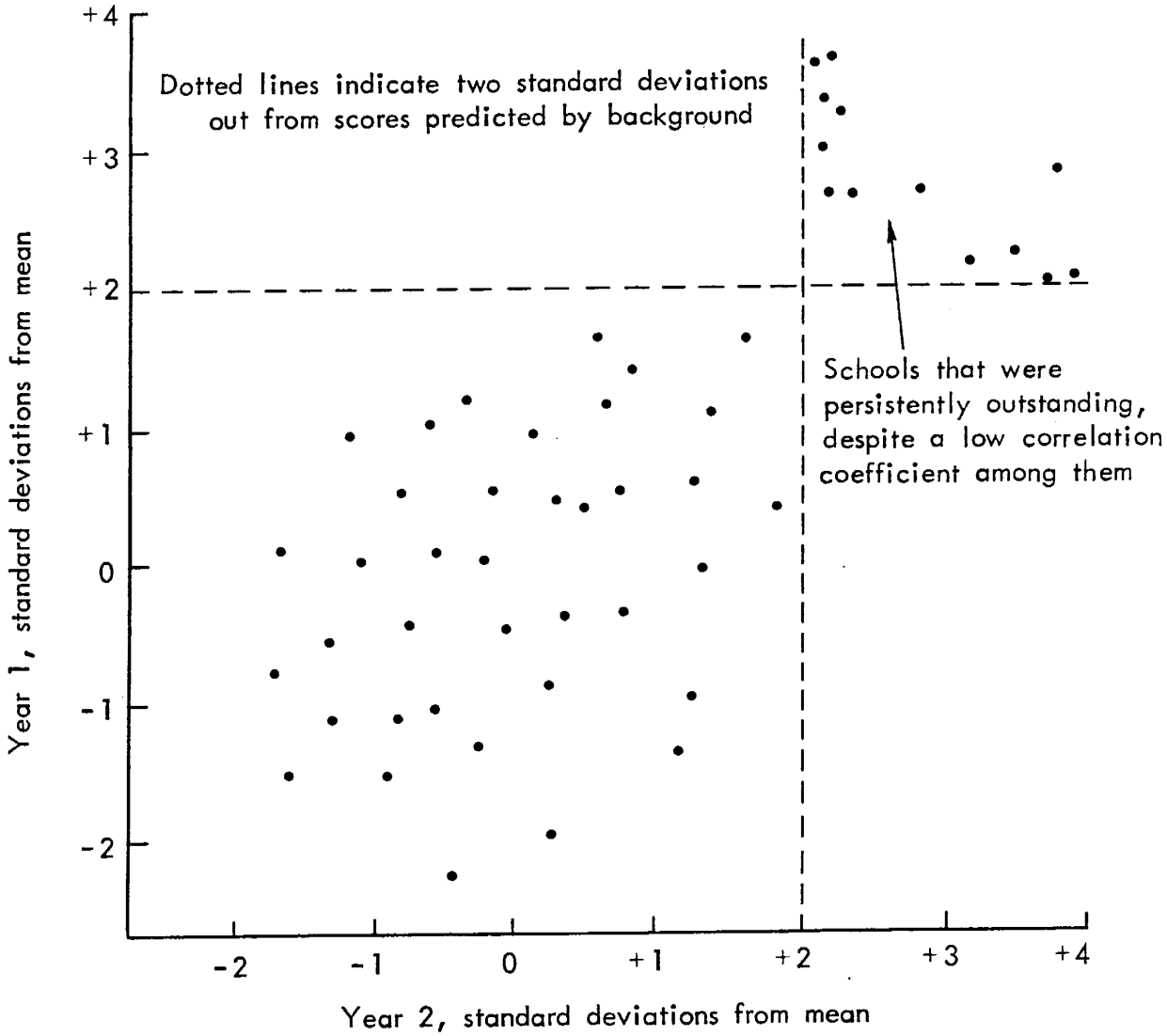


Fig.2 — Part of a hypothetical distribution of residuals

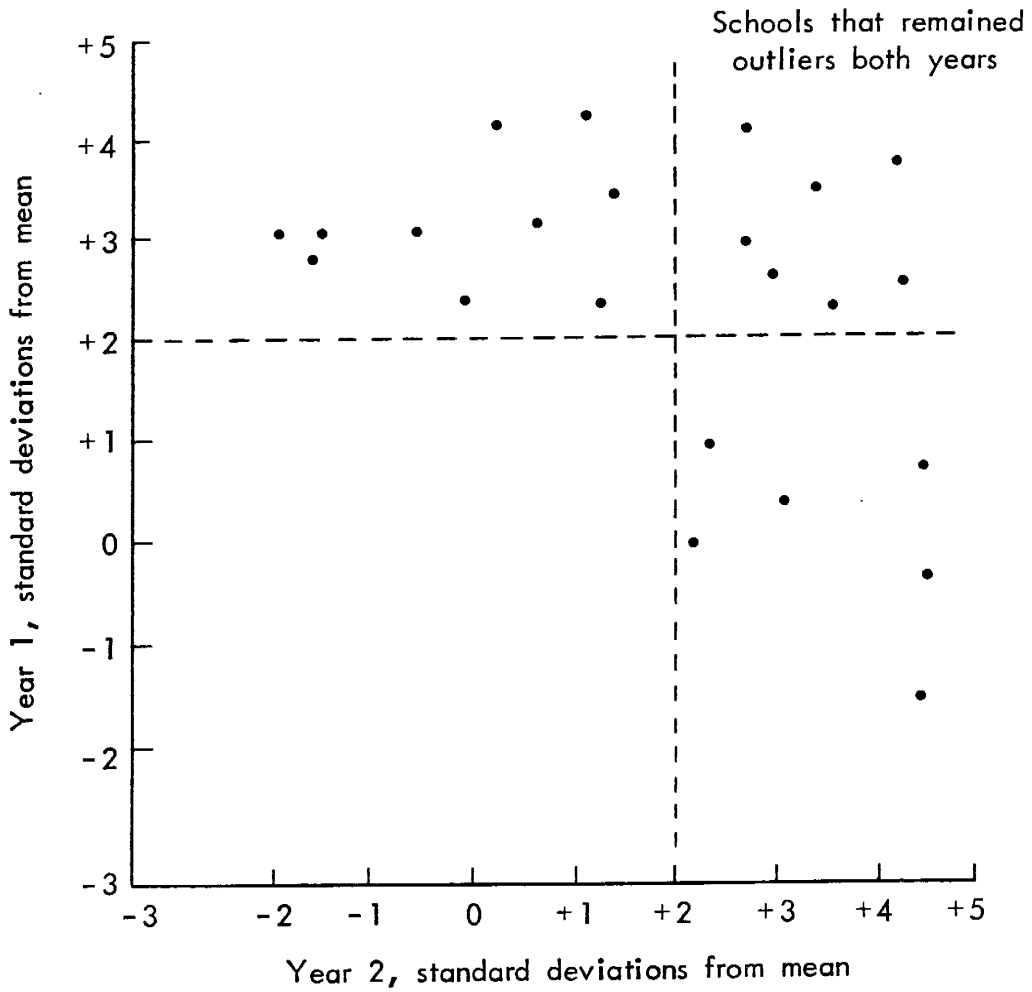


Fig.3 — Part of a hypothetical distribution of residuals showing persistence of outliers

The relevant question, then, is not the value of the correlation coefficient, but whether more schools remain outliers than chance alone would predict. Despite the thorough nature and importance of their work, Jencks and his associates did not really answer this question.

One of Jencks' coauthors has extended the analysis in another recent work. Acland (unpublished) examined the effectiveness of the average New York City elementary school using longitudinal data. He concluded that

> little evidence can be adduced from these analyses to upset the view that the average performance of students in an elementary school is primarily a function of the quality of the class when it entered the school. (P. 10.)

However, he too did not explore the performance of unusually effective individual schools. His conclusion, therefore, may or may not apply to outliers.

Acland's data forms a matrix with some 600 schools arrayed vertically and six grades (first through sixth) horizontally. For each cell (school and grade), Acland obtained the mean correlation over all school achievement score averages. The correlation coefficient between pairs of raw column means is very high, near 0.9, which parallels the Project Talent findings for high schools.

Acland also analyzed a residual measure of school effect. The 0.9 correlation between achievement scores in successive grades means one can accurately predict Y_{ij} by knowing $Y_{i,j-1}$, where Y_{ij} is the achievement score for school i in year j. Using a simple regression model,

$$\hat{Y}_{ij} = a + bY_{i,j-1} ,$$

where a and b are constants, Acland computed a residual $(Y_{ij} - \hat{Y}_{ij})$ for each cell. A positive residual for a cell may indicate that a school is contributing more than average to its students' scores in a particular grade. If a school contributes more over all grades, Acland reasoned, its positive residuals would be highly correlated.

Acland examined the correlation among the residuals in individual schools and found the mean correlation over all schools to be about 0.1, quite low. Thus, elementary schools exhibited no average tendency to consistently raise or lower the achievement scores from those predicted by the simple regression model. But, like Jencks, Acland aggregated over all schools; he does not discuss the performance of individual schools. Thus, his analysis does not deal with the existence of unusually effective performers.

Another, rather neglected analysis of exceptional schools merits attention. In 1968, the Office of Education reanalyzed the EEOS data and compared the residuals of over- and underachieving elementary schools. Figure 4 shows a histogram of the residuals from the basic regression between SES variables and achievement scores for the EEOS data. A small but interesting abnormality appears in the right tail.

To determine whether this was just a chance deviation, the authors compared the top and bottom 100 schools, ranked according to their residuals. These two groups of schools were found to differ significantly with respect to many school and situational characteristics. Among other differences, overachieving schools tended to have more parental interest, more and better instructional equipment, smaller class size, 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.

This paper contains two important results. First, the variables always considered important by educators were found to distinguish between underachieving and overachieving schools, although these variables did not account for much variation over all the schools in the Coleman report. Second, the significant differences in variables besides achievement are powerful evidence that the position of the top-ranking schools was not a mere statistical artifact.

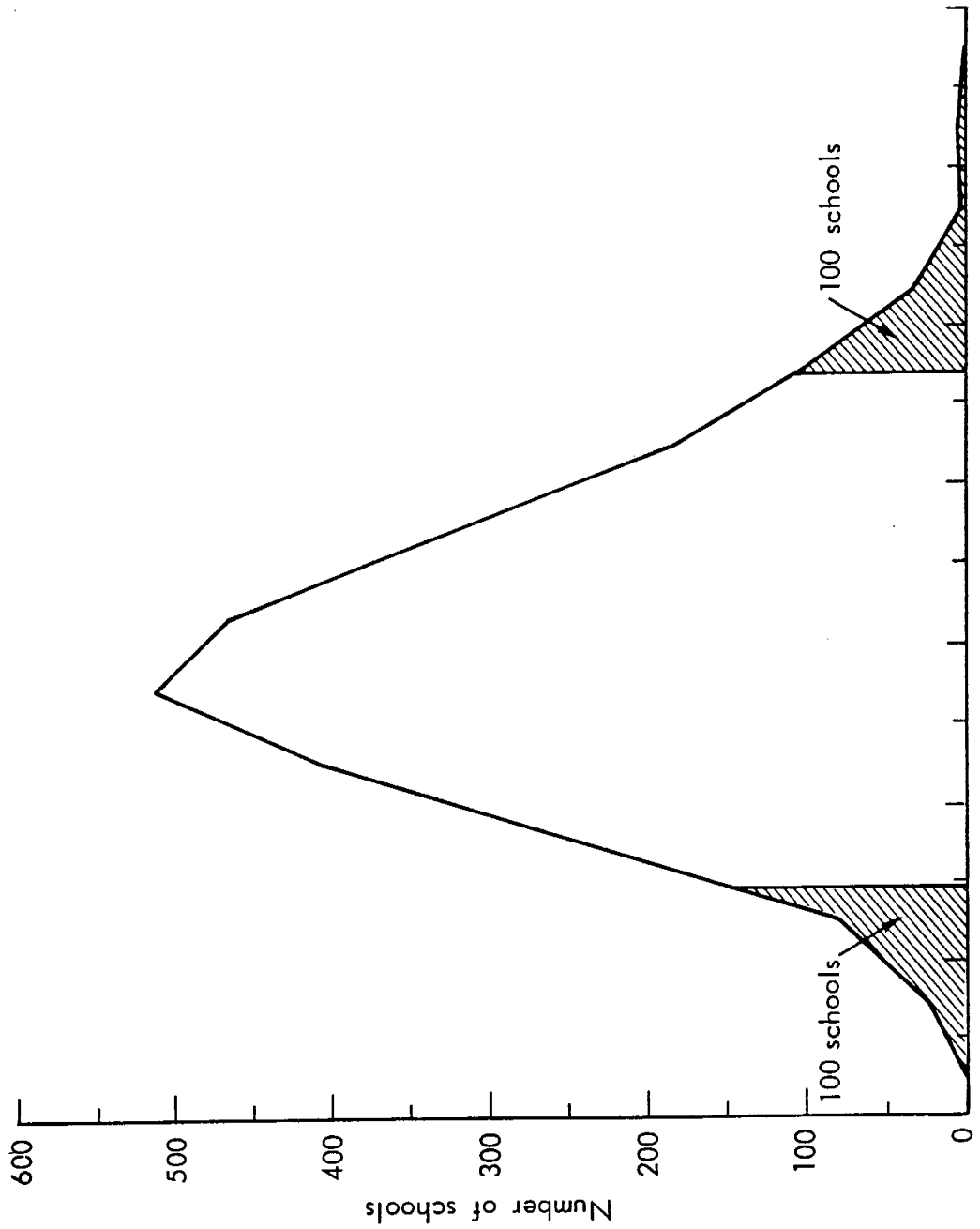


Fig.4 —Histogram of residuals from regression between SES variables and achievement scores for the EEOS data

III. STATISTICAL ANALYSIS

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 use 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 pay special attention to the *residuals* from the regression line.
- Instead of explicitly including school variables in the regression equation, we control for only socioeconomic and other nonschool background factors that affect achievement, and assume that the variation remaining 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 try to avoid overcontrolling.

Three reasons dictate these departures from previous practice. First, previous studies have shown that educational achievement is determined largely by nonschool factors. That is, both school effects and purely random fluctuation have been small. Residual variation could arise from many causes besides school differences: imperfections in measurement, misspecification of background factors, omitted variables, poor choice of fitting technique, incomplete data, and the combined random fluctuations involved in all the regressor variables.

Previous studies, however, by dint of their high R^2s, imply that such errors are not likely to be large. Such findings do not mean, as we shall see, that we can attribute residual effects solely to schools; but from past experience we can take comfort in the expectation that systematic errors will be small.

Second, statistical problems hinder the identification of a school's unique effect. If school and nonschool variables suffer from multicollinearity (Bowles and Levin 1968) or somehow have a joint effect not attributable to school or background alone (Mayeske, et al. 1969), judging the true impact of schools becomes nearly impossible. One might reason that, since we are seeking outstanding schools that are replicable, we should perhaps run two-stage least-squares regressions or specifically include an interaction term in the regression. That way, we would not classify as a school effect any variable that was inextricably bound up with socioeconomic and other background factors. But this argument is inappropriate here. We do not want to prejudge the question of replicability, nor eliminate school effects that are intercorrelated with background effects. Most important, *there is no convincing model showing the school variables that should be included explicitly 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.

Third, our null hypothesis posits that no unusually effective schools exist; if true, this hypothesis has serious consequences for educational policy. The significance of affirming the null hypothesis requires us to be very sure we do not accept it if it is false (we want to avoid a Type II error). Controlling for too many background variables increases the chance that, because of statistical interactions, real outliers will not show up. Controlling for too *few* variables risks the identification of "outliers" that could be explained by some missing regressor. However, the absence of outliers under such circumstances would be a strong result indeed. The best strategy is thus to avoid the risks of overcontrolling and to give exceptional schools every opportunity to exhibit themselves by regarding the entire residual as the school's effect, even though this practice imparts an upward bias to the estimate.

The outliers displayed by our method may not be due to unusually effective schools. They may be merely the product of chance perturbations or various statistical errors. But our task may be likened to that of a detective, in contrast to that of a judge. The detective searches for clues, the judge evaluates them. Our task is to find *prima facie* evidence for unusually effective schools, not to prove their existence absolutely. Only after candidates for exceptional schools are studied thoroughly can the verdict come in.*

Basically, the task is to find school outliers on achievement scores that are not explained by nonschool factors or random variation. Histograms of the residuals on socioeconomic factors from a regression of school achievement scores, as in Fig. 5, provide a good starting point. Histograms allow easy visual inspection for "lumpiness" in the distribution or unusual right tails, both of which may signal the presence of exceptional schools. Lumps would show that groups of schools are massed together in a discontinuous fashion, which may indicate several schools using different educational technologies or procedures. 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 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.

Advocates of the confirmatory approach to statistical analysis may be uneasy with such visual analyses and intuitive syllogisms. First, no theorem states that the residuals from a "good" fit should be normally distributed, since data seldom permit assumptions about the distributions from which the regression residuals were obtained. Second, attribution is difficult. Nonnormality would show up first in "unusual" tails; however, even if there were no "outstanding" schools, but just a nonnormal distribution of one of the variables, the tails might be oddly shaped. Third, in a confirmatory analysis, the tails must be

---

*We adopt here the philosophy of John Tukey. For his use of the judge-detective metaphor, see Tukey (1970), Vol. 1, pp. 1-2.

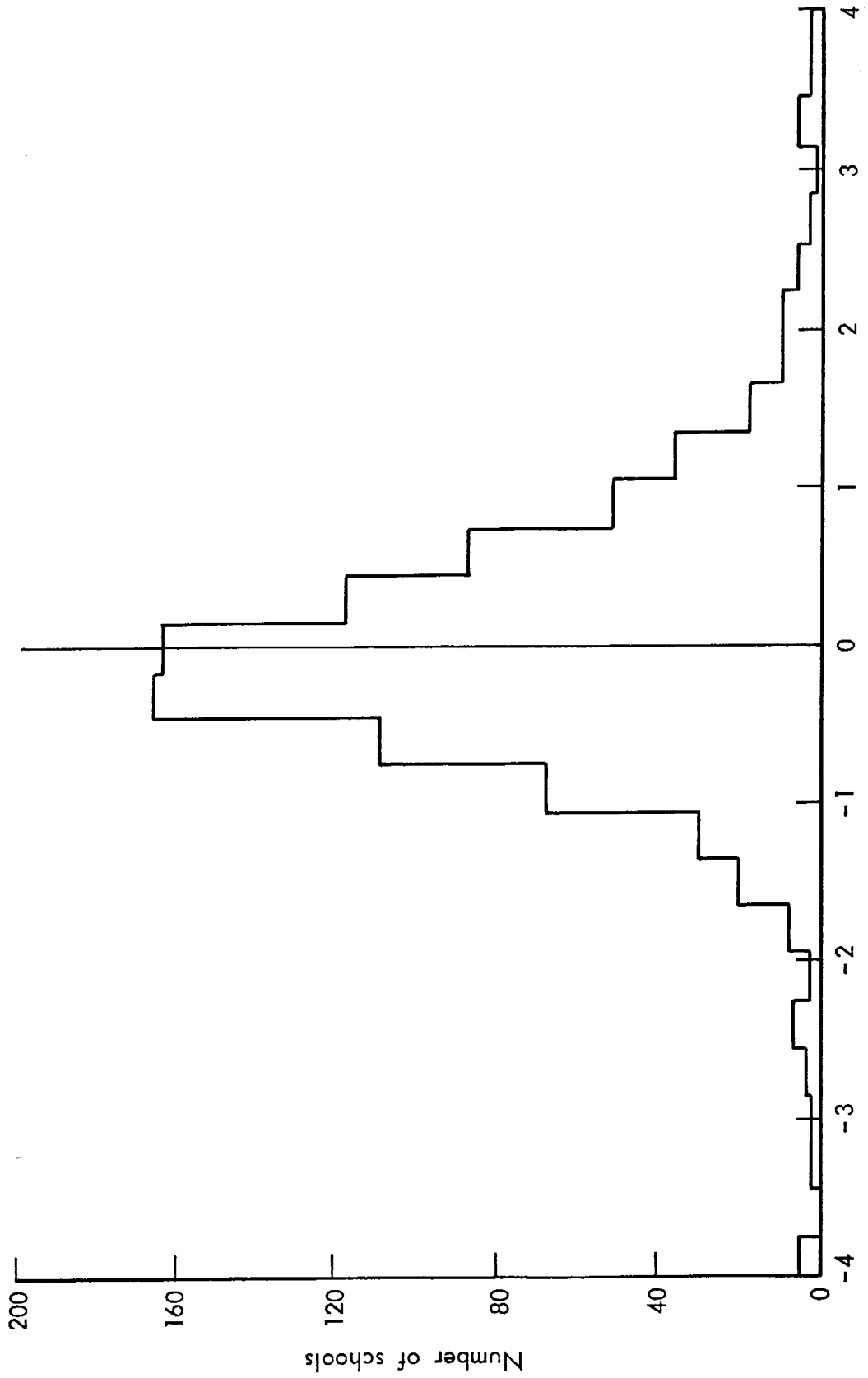


Fig.5 — Example of a histogram of residuals

very unusual to indicate that the distribution is not normal. When estimating confidence intervals around a fitted line, one can be most certain near the mean of the data; the confidence interval widens as one gets farther and farther from the mean. Such ambiguity makes it very difficult to say, using a fluctuation in the tails, that a given distribution is *not* normal. And since the residuals need not be normally distributed--that was the first point above--the task of proving that the tails are unusual becomes even more difficult.

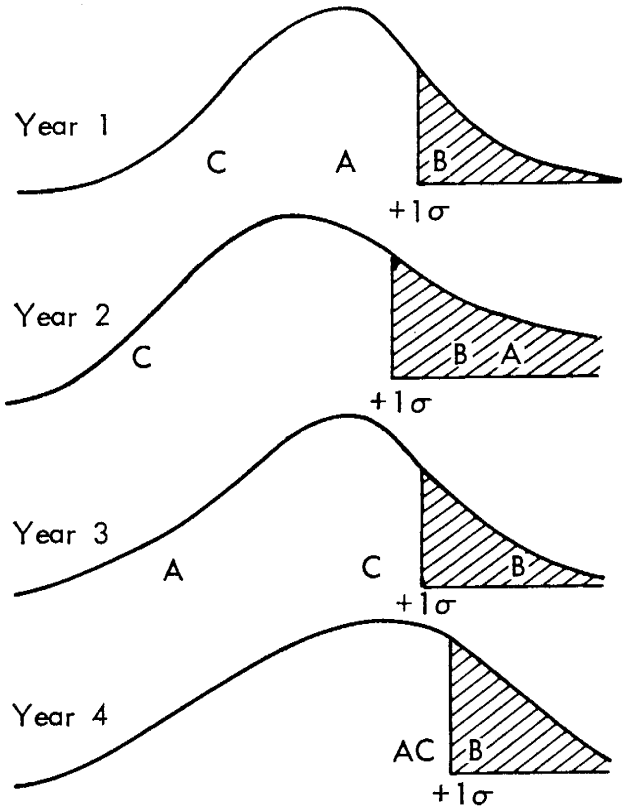
However, we are not attempting to prove absolutely that the tails are unusual; we are only exploring for evidence that warrants further investigation. Neither lumpiness nor unusual right tails would constitute conclusive evidence of anything; but they would provide interesting clues of where to concentrate our attention.

The second tool involves looking at *series* of distributions of residuals. Each individual distribution (e.g., 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 further study.

The null hypothesis says that all schools are equally effective, that all the variation in a particular distribution of residuals results from chance. To test this hypothesis, we would like a cumulative measure of school performance over many distributions, after controlling for background factors, to see whether the distribution of cumulative measures differed significantly from a theoretical distribution obtained by treating all the individual distributions of residuals as statistically independent.

To create this cumulative measure, each school in a given distribution (for a particular year, say) was assigned a score of one if it was more than one standard deviation above the mean and a zero otherwise. Then each school's totals were added up over all years considered, and we searched for schools consistently above one standard deviation.

For example, assume a set of data for schools for a certain grade during four successive years. The calculations of the cumulative measure are indicated in steps 1 and 2 of Fig. 6. Step 3 computes the theoretical



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 > 1σ.

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.

$$p(x \text{ successes}) = \binom{4}{x}(0.16)^{x}(0.84)^{4-x}$$

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

Number	Theoretical		Observed
0	P=0.50	404	400
1	P=0.38	307	301
2	P=0.11	89	95
3	P=0.014	11	10
4	P=0.0007	1	2

Note: $\chi^2 = 4.370$ (not significant)
Degrees of freedom = 4

Fig.6 — Hypothetical illustration of analysis of sets of residuals

distribution, using the binomial theorem and, in this case, a (constant) probability of 0.16 that a school would be more than one standard deviation above the mean in any one distribution. Step 4 compares the actual and expected distributions, using the Chi-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 appear to be outliers, it is important to see how they differ from the average school. Since in this report we are trying to discover only *if* such schools exist and not *why*, the point of the comparison is not to uncover causal mechanisms, although we may find some clues. Our goal is to separate nonrandom outliers from random variations. If many school-related characteristics of the top performers differ from those of the average school, they will provide strong evidence that we have indeed located schools worthy of detailed study, and not mere statistical quirks. On the other hand, if the only differences are in socioeconomic factors, the outliers may be the result of an omitted variable or heteroskedasticity in the regression.*

MICHIGAN SCHOOL DATA

The largest of the six data sets examined in this study is the school file compiled by the state of Michigan for 1969-1970 and 1970-1971. More than 2400 schools reported fourth-grade test results for both years, and some 900 schools had seventh-grade results for both years. Most of the schools with seventh-grade results also provided fourth-grade scores. In addition to large sample size, the data have other desirable properties--timeliness, relative novelty,† and a large range of variables.

---

*Heteroskedasticity refers to nonconstant variance of residuals around the regression line. If, for example, schools in wealthy areas show greater variability in their achievement scores than those in low and middle income areas, most of the high (and low) outliers will be the wealthier schools. The problem is distinguishing "real" socioeconomic differences in residual variation from those occasioned by some property of the socioeconomic measure. This difficulty arose with respect to rural schools in the Michigan data, as discussed below (pp. 24, 40-43).

†Another analysis of the Michigan fourth-grade data has recently appeared (Brown 1972).

Schools' average scores for five basic achievement tests and a composite were reported, as well as information on policy-related variables, such as staff size, pupil/staff ratios, and teacher characteristics, though these differed somewhat between the 1969-1970 and 1970-1971 data sets. Socioeconomic data came from questionnaires filled out by students when taking the achievement tests; three separate SES indicators and one composite SES index were computed by the state of Michigan for each school. In addition, other questions measured student "friendliness" and attitudes toward achievement, self, and school. Approximately 12 percent of the schools reported no achievement score data at all, so the results do not reflect a complete sample of Michigan schools.

Preliminary Explorations

We examined both mathematics and reading achievement as separate indicators of school effectiveness, and we computed sets of eight regressions on both tests for both grades for both years. To give outliers every opportunity to display themselves, in our first runs we controlled only for socioeconomic status, using the composite SES index for each grade and year. The results appear in Table 1.

Table 1

REGRESSIONS ON MICHIGAN DATA, 1969-1970 AND 1970-1971, FOURTH AND SEVENTH GRADES, READING AND MATHEMATICS TESTS

Test	Equation	F-ratio	R^2	Control
R769-70	Reading = 32.3 + 0.37(SES)	238	0.20	SES769-70
M769-70	Math = 32.0 + 0.38(SES)	187	0.17	SES769-70
R469-70	Reading = 28.3 + 0.44(SES)	750	0.23	SES469-70
M469-70	Math = 27.4 + 0.46(SES)	802	0.24	SES469-70
R770-71	Reading = 20.5 + 0.60(SES)	739	0.46	SES770-71
M770-71	Math = 16.1 + 0.69(SES)	723	0.45	SES770-71
R470-71	Reading = 20.7 + 0.60(SES)	1949	0.44	SES470-71
M470-71	Math = 18.5 + 0.64(SES)	2107	0.46	SES470-71

NOTE: R769-70 stands for the regression on reading scores for the seventh grades in 1969-1970. The other symbols are interpreted similarly.

The large range in the percentage of variation explained in the different regressions (from 0.17 to 0.46) was surprising. The regression coefficients, F-ratios, and R^2 statistics are all higher for 1970-1971 than for 1969-1970; however, differences between neither grades nor tests are systematic. We therefore suspected a change in the definition of one or more variables from one year to the next. Investigation of the SES questionnaires revealed that a definite shift had occurred between the two years. The 1970-1971 questionnaire was longer and omitted questions about parent mobility that had apparently overestimated the SES for central-city areas and underestimated the SES of rural districts.*

Both the reading and mathematics achievement scores are standardized nationwide to have a mean of 50 and an interstudent standard deviation of 10. The range of school mean scores uncontrolled for background factors is large. Between the highest and lowest schools, the range is about 35 points; the difference between the mean scores for the top and bottom 100 schools is around 13 points; and the standard deviation among schools is about 4 points.†

The histograms of the residuals are instructive primarily for what they do not show. The overwhelming impression is how normal everything is. Two typical histograms are displayed in Figs. 7 and 8.‡ The residuals all display unimodal massing around the mean with little skewness. The small amount of skewness in evidence is inconsistent: skewness statistics range from -0.94 to 0.35; five are negative, three positive. Some tails wander out to four standard deviations, but, given the large sample size, such extensions are not surprising. No lumps or

---

*Mobility is usually associated with increased affluence. It is well known, however, that central-city residents are often highly mobile but poor, while the rural population tends to be more stable than their other SES characteristics would seem to warrant. As a result of the shift in the Michigan questionnaires, the mean SES of rural Michigan communities "rose" from 47.5 in 1969-1970 to 49.2 in 1970-1971.

†All these figures were computed from the 1970-1971 fourth-grade reading data.

‡We plotted nearly 40 additional histograms, all very much like Figs. 7 and 8.

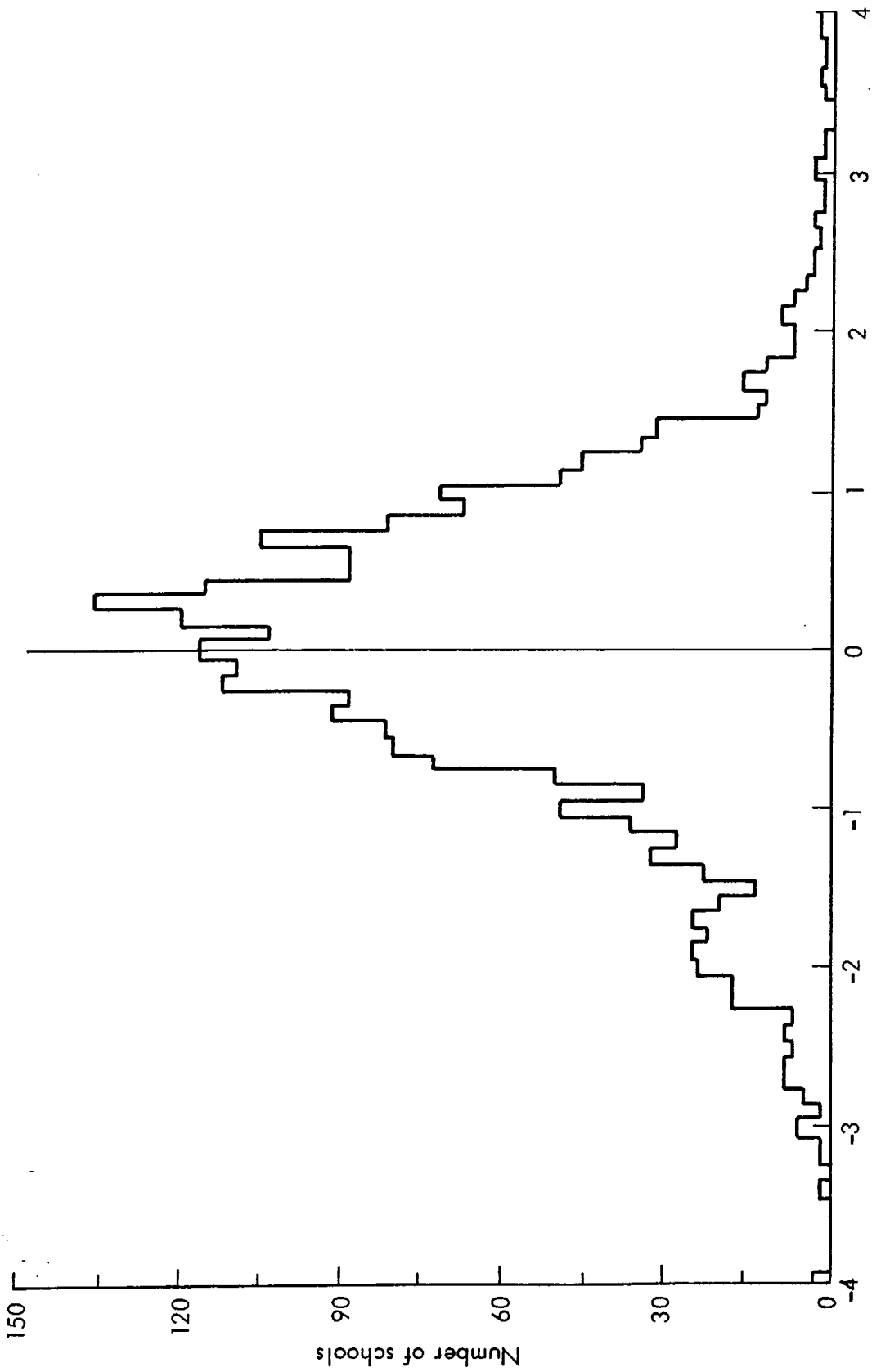


Fig. 7 — Histogram of residuals for 1970-1971 Michigan fourth-grade mathematics test, after controlling for socioeconomic status

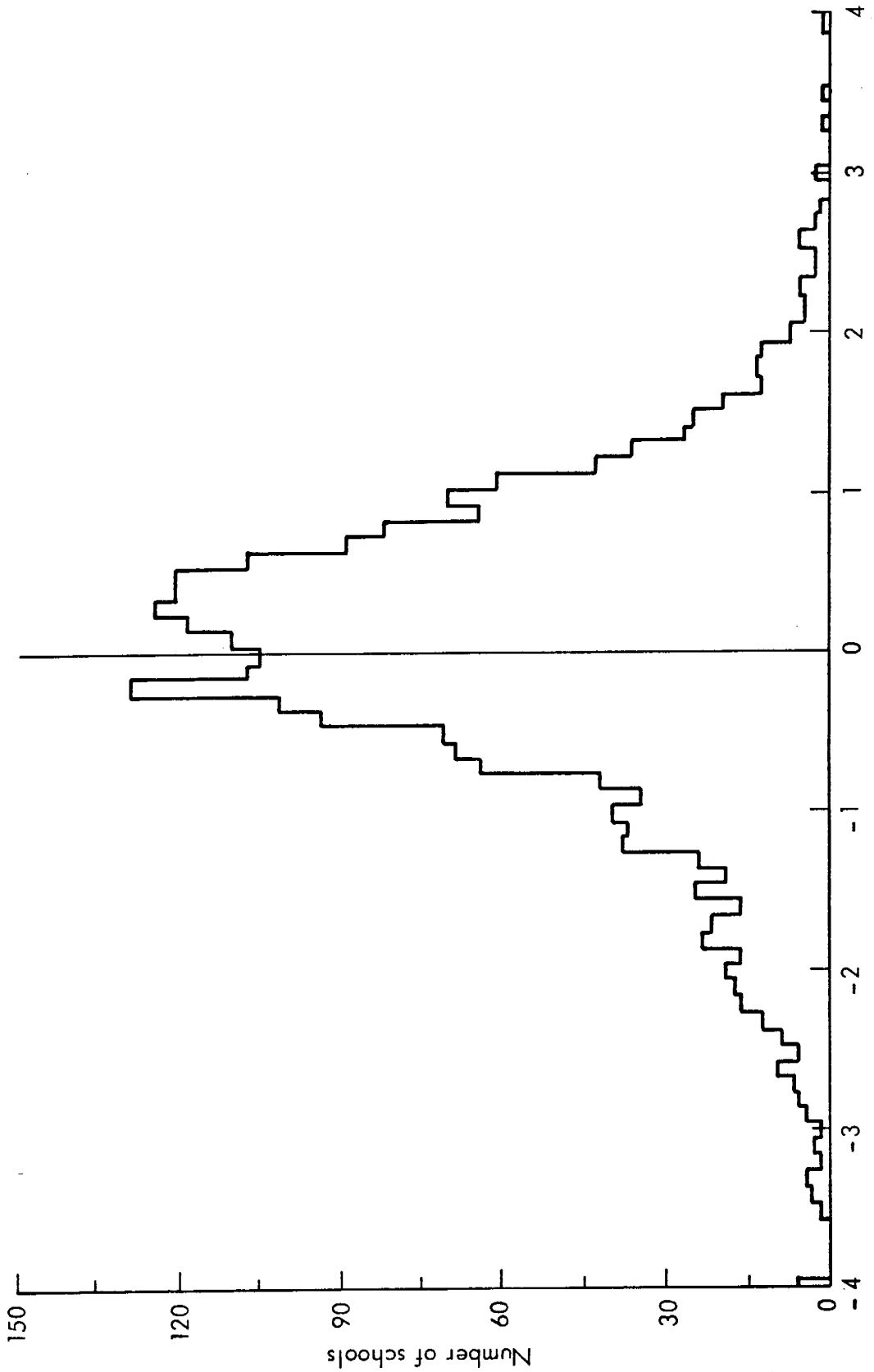


Fig. 8 — Histogram of residuals for 1970-1971 Michigan fourth-grade reading test, after controlling for socioeconomic status

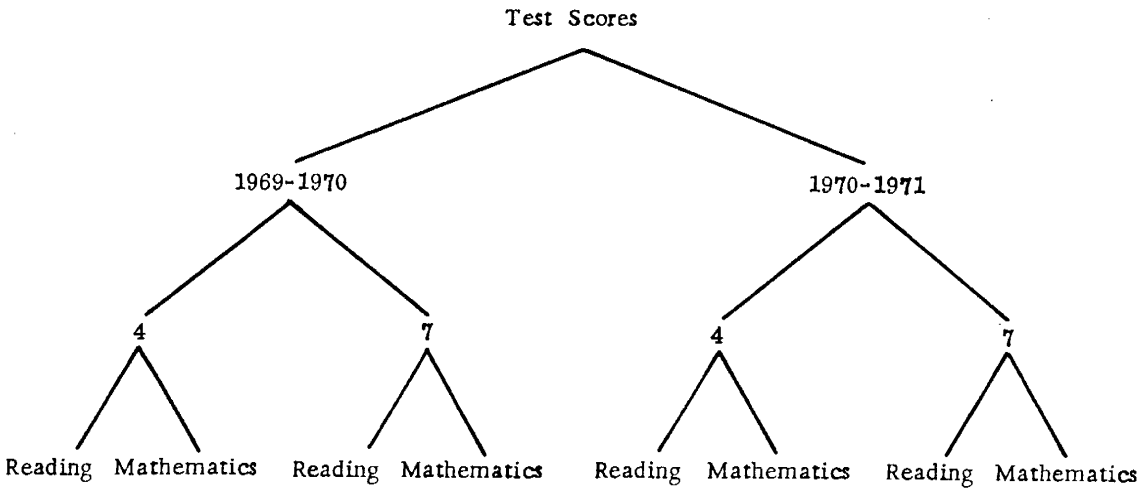
discontinuities appear in the distributions. In short, the histograms reveal a most well-behaved set of residuals--especially since only SES was used as a control. There is no hint that some schools are extraordinarily far above the mean nor that there is discontinuity in educational effectiveness among schools.

However, some schools may remain in the right tails of the distributions of residuals over time, even if each distribution appears to be normal. We checked to see if some schools consistently had residuals more than one standard deviation above the mean, using the cumulative measure outlined earlier (pp. 20-22). There were three complications to this analysis. First, not every school had a score for each test in each year and grade. (For example, less than 10 percent of the schools appeared in all eight regressions.) Schools were therefore grouped by how many times out of eight "chances" they reported scores. Different theoretical distributions were computed for schools with eight, six, four, and two chances to be more than one standard deviation out.

A second problem involved the correlation of the mathematics and reading residuals. The correlation between the raw scores for mathematics and reading in the same grade and year was very high, 0.9; and the residuals were correlated about 0.78. This high correlation is not surprising, since the same children were taking the tests, but we had to modify our assumptions that the different residuals were strictly independent.

The following tree shows the eight residuals. There are two years, two grades per year, and two tests per grade. Only the last four pairs of forks are not assumed to be independent by the null hypothesis that no consistently exceptional schools exist. Since the reading and mathematics residuals were correlated about 0.78 for each year, the null hypothesis might be reworded: the four *pairs* of test scores are independent, and each pair is bivariate normally distributed with $\rho = 0.78$. A theoretical distribution could then be computed and compared with the actual distribution using a Chi-square test for goodness of fit.

However, a third problem arose. The empirical distributions of the pairs of scores taken individually did not look bivariate normal. Let X_i be the number of scores in the reading-mathematics pair (R_i, M_i)



that exceeds one standard deviation above the mean. X_i has the possible values 0, 1, or 2. If a pair of scores is distributed bivariate normally with $\rho = 0.78$, then $P(X = 0) = 0.777$, $P(X = 1) = 0.128$, and $P(X = 2) = 0.095$. But the average pair of Michigan reading and mathematics scores actually displayed a different distribution: $P(X = 0) \cong 0.846$, $P(X = 1) \cong 0.088$, and $P(X = 2) \cong 0.066$.*

The actual distribution is "tighter" than the bivariate normal; that is, schools are less likely to be one standard deviation above the mean at least once in two chances. Also, $P(X = 2)$ is larger compared to $P(X = 1)$ than in the bivariate normal case, indicating that schools have a greater tendency to be outside both times, given that they are outside at least once.

A score T was computed for each school, where T was the sum of the X_i's for each of the school's n pairs of scores, i.e., $T = \sum_{i=1}^{n} X_i$. (In the case of a school reporting all four pairs of scores, $T = X_1 + X_2 + X_3 + X_4$.) Assuming the X_i's were independent, null distributions were

*The actual probabilities for each pair of tests were:

Grade/Year	N	P(X=0)	P(X=1)	P(X=2)
469-70	2485	0.841	0.091	0.068
769-70	942	0.868	0.076	0.055
470-71	2469	0.833	0.097	0.070
770-71	876	0.871	0.064	0.065

computed using the actual probabilities of 0, 1, and 2 successes per pair, instead of a bivariate normal with $\rho = 0.78$. Then the actual distributions were compared to the null distributions using a Chi-square test.*

Table 2 presents the results of our search for schools that consistently had residuals more than one standard deviation above the mean. Here is strong evidence that some schools are persistent outliers. In every case, there are more schools with residuals greater than one standard deviation above the mean than would have been expected on the basis of chance alone. For example, among schools with reported scores for each test/grade/year regression, chance would predict that only five would have four or more residuals (out of eight chances) greater than one standard deviation above the mean, whereas in fact 34 did.†

However, significant Chi-square statistics alone must not be taken as evidence of consistent positive outliers, since the Chi-square statistic reflects differences between the actual and the estimated distributions throughout the entire distribution. A larger number of underachievers than expected can therefore affect the Chi-square as much as a larger number of overachievers. The Chi-square statistics in Table 2 are therefore less important than the visual evidence that consistent overachievers may exist.‡

---

*If a school had eight chances to be outside, the null hypothesis worked from a basis of four pairs of tests; if only six chances, then three pairs of tests; if four chances, two pairs. The chances occurred only in reading-mathematics pairs (any school that reported a reading score for a given grade and year also reported a mathematics score for that grade and year). For simplicity in calculation, we computed the probability for an average Michigan pair from all grades and years and based our null distributions on it. This procedure assumed it did not matter which particular test pairs made up a school's set of chances, which seemed reasonable in light of the relatively constant sizes of the probabilities across different grades and years.

†In order that the Chi-square approximation be accurate in contingency tables with more than one degree of freedom, one must pool cells with small expectations. Throughout this report, we follow a pooling rule proposed by Yarnold (1970, p. 865):

> 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 $5r/s$.

‡Of course, if the Chi-square statistic is not significant, it provides strong evidence that consistent outliers do not exist.

Table 2

RESULTS OF CHI-SQUARE TESTS OF DIFFERENCES BETWEEN OBSERVED AND EXPECTED DISTRIBUTIONS OF RESIDUALS FOR MICHIGAN SCHOOLS

Schools Reporting 8 Times			Schools Reporting 6 Times			Schools Reporting 4 Times		
No. >1	Observed	Expected	No. >1	Observed	Expected	No. >1	Observed	Expected
0	69	84	0	90	102	0	1995	1893
1	31	34	1	37	31	1	309	383
2	18	30	2	22	27	2	197	307
3	9	8	3	6	5	3	69	32
4	10		4	8		4	56	11
5	6		5	4	2			
6	8 (34)	5	6	0 (12)				
7	3							
8	7							
$\chi^2 = 200.9$, Degrees of Freedom = 4			$\chi^2 = 46.7$, Degrees of Freedom = 4			$\chi^2 = 301.1$, Degrees of Freedom = 4		

NOTE: All residuals are derived from a fit of achievement to SES alone. All Chi-square statistics are significant at the 0.005 level.

Since we controlled only for SES in determining these residuals, the outliers could reflect some other, uncontrolled background variables and not the existence of overachieving schools. To examine this possibility, we first ranked schools on the basis of one of the regressions (the size of their fourth-grade reading residuals for 1970-1971) and then compared 19 school and background variables among the top 100, middle 100, and bottom 100 schools. The results appear in Table 3. Our goal was to determine whether the top schools differed significantly from the others in terms of some omitted nonschool factor.

Table 3 shows that background variables such as the racial composition of the schools and community type sharply differentiate the three groups of schools. The bottom 100 are 38 percent white; the top 100 are 90 percent white, even after controlling for SES. The rural/urban index also shows large differences. The more rural a school, the more likely it is to have larger residuals. These results suggest that further regressions should be run that explicitly control for race and community type. Otherwise, the residuals will reflect large differences in background factors instead of differences in school effectiveness.

Table 3

COMPARISONS AMONG THE TOP 100, MIDDLE 100, AND BOTTOM 100 MICHIGAN SCHOOLS RANKED BY RESIDUALS ON THE FOURTH-GRADE READING TEST, 1970-1971

Variable	Bottom 100		Middle 100		Top 100		*t*-statistics		
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	B-M	M-T	B-T
Teacher characteristics									
Pupil/teacher ratio	25.1	0.4	25.6	0.4	24.4	0.4	-0.8	3.0*	1.2
% Teachers with over 5 years' experience	57	17	55	13	65	21	0.8	-7.5*	-4.8*
% Teachers with master's degrees	27	15	26	16	20	18	0.2	3.6*	4.2*
% Teachers earning over $11,000	41	20	38	22	27	26	1.0	5.2*	7.3*
Test scores									
Vocabulary	43.3	3.8	51.1	3.4	54.4	4.4	-15.4*	-9.8*	-29.3*
Reading	41.3	3.3	51.5	2.5	55.8	2.4	-24.5*	-16.7*	-43.5*
Mechanical engineering	42.4	3.9	51.3	2.9	54.5	3.9	-18.2*	-11.4*	-30.8*
Mathematics	42.6	4.3	51.5	3.2	55.0	4.0	-16.7*	-10.9*	-29.0*
Composite	42.2	3.6	51.5	2.7	55.1	3.0	-20.5*	-13.3*	-35.5*
Attitude scores									
Toward achievement	51.6	3.5	50.1	2.3	49.9	4.6	1.2	5.3*	4.9*
Toward self	50.3	2.4	49.9	2.6	48.8	4.4	0.9	4.5*	6.1*
Toward school	49.3	3.3	49.9	2.3	50.8	3.8	-1.6	-4.1*	-4.8*
"Friendliness"	49.3	5.1	50.5	2.6	50.8	5.0	-7.2*	-1.3	-8.8*
Background factors									
% White enrollment	38	38	89	15	90	12	-12.5*	-0.6	-13.7*
Composite SES (regressor)	46.6	5.0	51.3	4.2	47.2	4.1	-7.3*	9.6	-1.3
SES factor 1	45.2	5.3	51.3	3.1	50.1	5.2	-10.0*	3.8*	-9.3*
SES factor 2	50.2	3.6	50.4	4.3	46.5	4.1	-0.4	9.0*	10.0*
SES factor 4	47.4	3.0	50.1	2.8	49.2	4.9	-6.7*	3.5*	-5.9*
Urban/rural (urban core = 1, rural = 5)	2.1	1.6	3.3	1.4	3.8	1.4	-5.6*	-3.5*	-10.6*

NOTE: An asterisk (*) indicates difference not equal to zero at the 0.005 significance level.

Table 3 reveals interesting results along school variables as well. The pupil/teacher ratio does not effectively differentiate among the three groups, although the top schools had slightly smaller class sizes. The rural-urban split may explain some surprising results regarding teacher characteristics. The bottom 100 were more likely to have teachers receiving over $11,000 a year and with master's degrees, but less likely to have teachers with over five years' experience. Since salary usually increases with seniority, this result seems strange. But rural areas may pay teachers less and have fewer teachers with master's degrees, yet suffer less turnover than urban areas. Rural-urban differences may be reflected in the labor market for teachers, as well as in achievement scores.

Two other findings in Table 3 should be noted. First, all tests showed similar differences among the three groups of schools. A school that did well in one subject area also tended to do well in others. Second, the four personality tests indicated that students in the top schools had significantly higher scores on "friendliness" and attitude toward school, but, perhaps surprisingly, significantly lower scores on attitudes toward achievement and self.

Controlling for Additional Variables

As a result of Table 3, we ran new regressions that included controls for racial composition and community type.* Two other studies of the Michigan data also use community and regional differences as regressor variables. Brown (1972) found achievement differences that appeared to be related to rural-urban location. Arthur Alexander, in an unpublished investigation, compared 22 high-scoring schools with 32 low-scoring schools. Both sets of schools were also compared with the total group of 2400 schools. After selecting his samples, Alexander

*The Michigan data set classifies schools according to their location in one of five "community types":

1. Metropolitan core.
2. City (population from 10,000-50,000).
3. Town (2,500-10,000).
4. Urban fringe.
5. Rural (less than 2,500).

compared pupil/teacher ratios, class size, SES variables, community type, and differences in the characteristics of school principals. He found that the only major difference between the two groups was geographical. The high-scoring schools were concentrated in small towns and rural areas, whereas the others tended to be in large cities and industrial suburbs.

We investigated the changes in the relationship between achievement and background variables when dummy variables for community type are added to the estimating equations, and when the racial composition of the school is controlled. Table 4 shows some of the results. The addition of these variables increases the explanatory power of the regression. A comparison of Table 4 with Table 1 shows that the R^2s are now about 10 to 30 percentage points higher. Moreover, the explanatory power of the 1969-1970 regressions now matches that of the 1970-1971 regressions. The inclusion of racial composition and community type apparently compensates for the imperfections in the 1969-1970 SES questionnaire.

If a regional variable is not included, some of the community types enter with significant coefficients. If both community type and region are included in our equations, the regional coefficients are significant but the community-type coefficients have low F-values.

Interestingly, the histograms of residuals from these more complicated regressions show some evidence of unusual right tails. Figures 9 and 10 give two examples. Compared with the histograms from the simpler fit, these show far more schools three and four standard deviations above the mean. Apparently, these outliers were masked earlier by the variability due to racial composition and community type. The two additional background controls reduced overall variability, allowing unusually effective schools to stand out more in the right end of the distribution. Table 5 compares the actual number of outliers to the number expected if the residuals were distributed normally. The mathematics scores in particular show many schools on the right tail.

These deviations from normality are, however, not astoundingly large. We cited earlier a number of reasons to expect deviations from

Table 4

MICHIGAN SCHOOL REGRESSIONS, CONTROLLING FOR RACIAL COMPOSITION, COMMUNITY TYPE, AND SOCIOECONOMIC STATUS

Test	Equation	R^2	Standard Error	Number of Schools
R469-70	Y = 32.17 - 0.08(% minority) + 0.40(SES 4) - 1.09(C1) + 0.08(C2) - 0.97(C4) + 0.07(C5) *904.3* *829.9* *25.0* *0.1* *27.8* *0.1*	0.53	2.86	2485
M469-70	Y = 31.39 - 0.08(% minority) + 0.41(SES 4) - 0.83(C1) - 0.16(C2) - 0.82(C4) + 0.06(C5) *867.8* *844.6* *13.7* *0.4* *19.0* *0.1*	0.52	2.94	2485
R769-70	Y = 34.07 - 0.08(% minority) + 0.35(SES 7) - 0.43(C1) - 0.94(C2) - 1.00(C4) + 0.35(C5) *261.0* *246.4* *1.4* *5.1* *10.2* *1.4*	0.48	2.74	942
M769-70	Y = 33.79 - 0.09(% minority) + 0.37(SES 7) - 1.19(C1) - 0.21(C2) - 1.30(C4) + 0.46(C5) *275.5* *215.7* *8.5* *0.2* *14.1* *2.0*	0.50	3.04	942
R470-71	Y = 28.98 - 0.06(% minority) + 0.46(SES 4) - 1.17(C1) + 0.17(C2) - 0.92(C4) - 0.27(C5) *429.7* *1097.9* *33.9* *0.6* *30.5* *2.3*	0.58	2.64	2469
M470-71	Y = 26.99 - 0.06(% minority) + 0.49(SES 4) - 1.00(C1) + 0.26(C2) - 0.69(C4) - 0.19(C5) *417.4* *1181.1* *11.7* *1.2* *15.6* *1.1*	0.58	2.78	2469
R770-71	Y = 29.06 - 0.06(% minority) + 0.45(SES 7) - 0.31(C1) + 0.21(C2) - 0.29(C4) - 0.45(C5) *134.5* *346.2* *0.9* *0.3* *1.1* *2.0*	0.55	2.38	876
M770-71	Y = 27.96 - 0.07(% minority) + 0.48(SES 7) - 1.53(C1) + 0.10(C2) - 0.70(C4) - 0.34(C5) *165.0* *222.6* *17.8* *0.1* *4.4* *1.4*	0.60	2.59	876

NOTE: C1, C2, C4, and C5 are dummies for community types: 1 = metropolitan core; 2 = city (10,000-50,000); 4 = urban fringe; 5 = rural (< 2500). R469-70 stands for the regression on reading scores for the fourth grades in 1969-1970. The other symbols are interpreted similarly. Figures in italics below the regression coefficients are the F-ratios (= t^2).

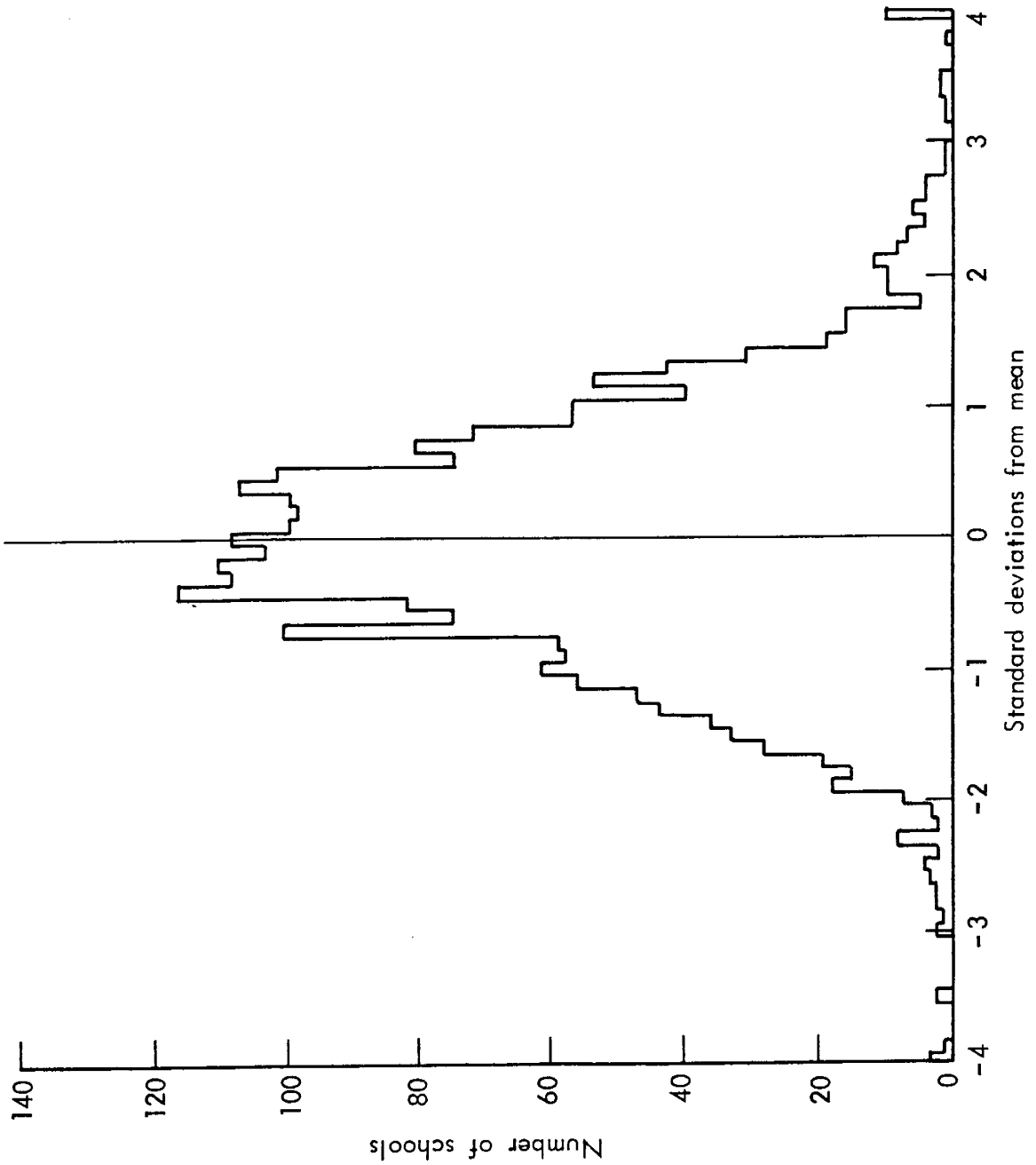


Fig. 9 — Histogram of residuals for 1969-1970 Michigan fourth-grade mathematics test, from a regression controlling for racial composition, community type, and socioeconomic status

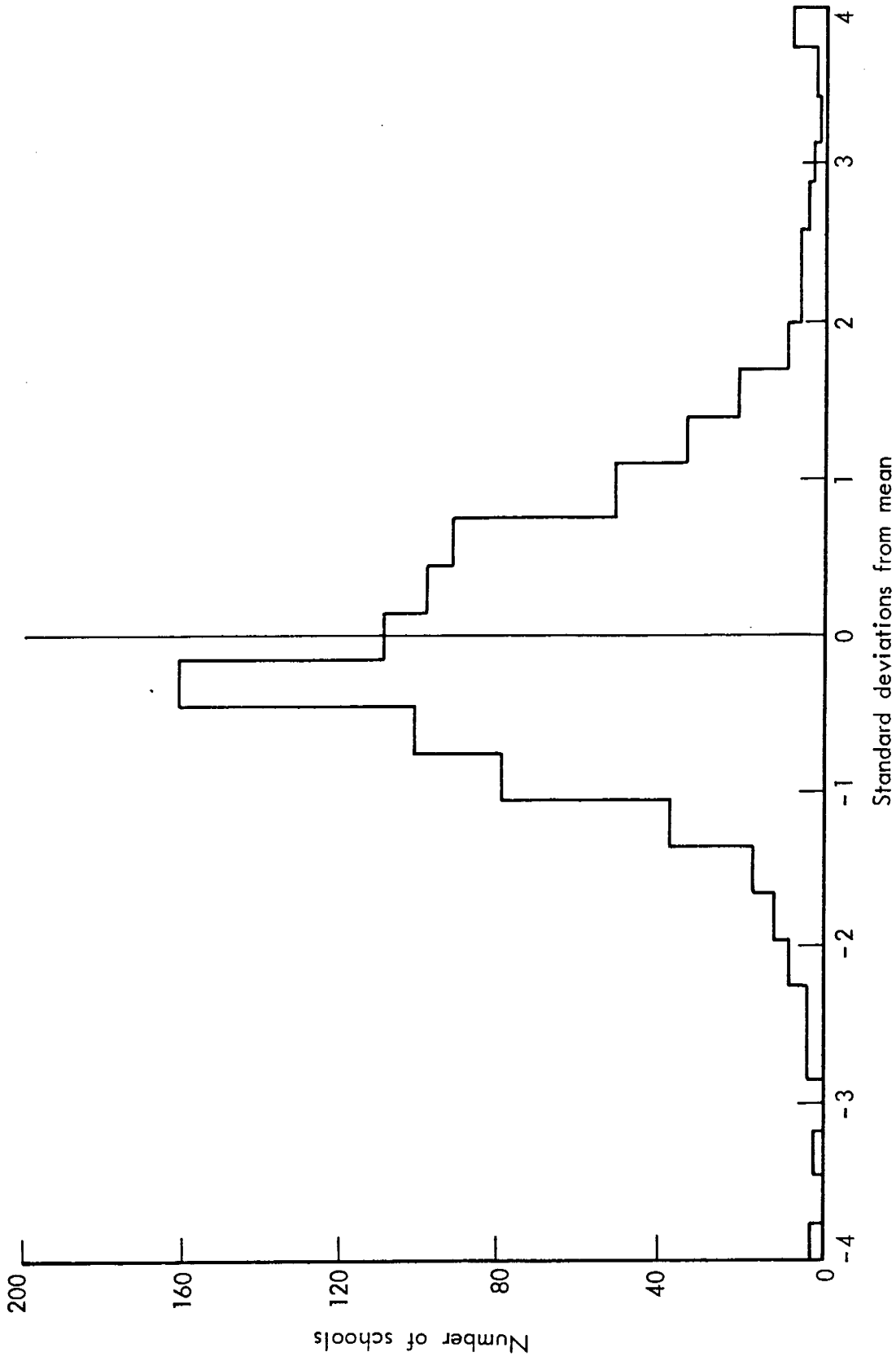


Fig. 10 — Histogram of residuals for 1970-1971 Michigan seventh-grade mathematics test, from a regression controlling for racial composition, community type, and socioeconomic status

Table 5

NUMBER OF EXPECTED AND OBSERVED MICHIGAN SCHOOLS WITH LARGE RESIDUAL TEST SCORES AFTER CONTROLLING FOR RACIAL COMPOSITION, COMMUNITY TYPE, AND SOCIOECONOMIC STATUS

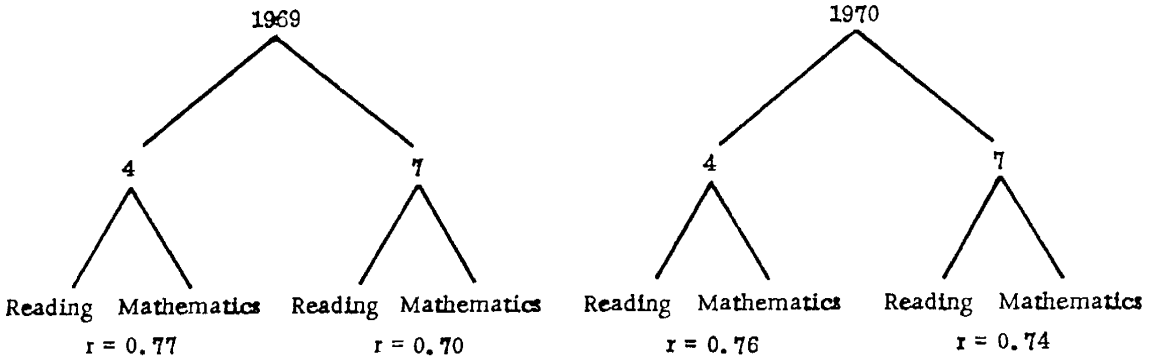
Residuals	Number Observed	Number Expected in Normal Distribution	Total Number of Schools
R469-70 ≥ 2 std. dev.	50	56.7	2486
≥ 3 std. dev.	8	3.2	
≥ 4 std. dev.	3	0.1	
M469-70 ≥ 2 std. dev.	75	56.7	2485
≥ 3 std. dev.	18	3.2	
≥ 4 std. dev.	10	0.1	
R769-70 ≥ 2 std. dev.	24	21.5	943
≥ 3 std. dev.	8	1.2	
≥ 4 std. dev.	2	0.05	
M769-70 ≥ 2 std. dev.	32	21.5	942
≥ 3 std. dev.	12	1.2	
≥ 4 std. dev.	1	0.05	
R470-71 ≥ 2 std. dev.	57	56.2	2469
≥ 3 std. dev.	16	3.2	
≥ 4 std. dev.	5	0.1	
M470-71 ≥ 2 std. dev.	79	56.2	2469
≥ 3 std. dev.	22	3.2	
≥ 4 std. dev.	7	0.1	
R770-71 ≥ 2 std. dev.	21	20.0	876
≥ 3 std. dev.	8	1.1	
≥ 4 std. dev.	3	0.05	
M770-71 ≥ 2 std. dev.	30	20.0	876
≥ 3 std. dev.	14	1.1	
≥ 4 std. dev.	4	0.05	

NOTE: R469-70 stands for fourth-grade reading scores in 1969-1970. Other symbols are interpreted similarly.

a normal distribution even if all schools had the same effectiveness (pp. 18-20). A more powerful test is available: as before, we can examine the distribution of residuals over tests and over time. When controlling for SES alone, we found that more schools than chance would allow consistently exhibited residuals greater than one standard deviation above the mean. If, after controlling for racial composition and community type, these outliers disappear, it would seem that our previous results were a function not of unusually effective schools but of previously uncontrolled background factors.

However, Table 6 indicates that, even after controlling for race and community type, there are still more outliers than chance would predict.* In fact, there are nearly as many schools above one standard deviation with the more complicated controls as with SES alone. The interschool standard deviation is now slightly smaller: the standard errors for the regressions using SES, racial composition, and community type range between 2.38 and 3.04, as opposed to a spread of 2.61 to 3.92 with SES alone. The new results therefore suggest

*The reading and mathematics residuals were again correlated for each grade and year, in the following fashion.



The empirical probabilities for each pair are:

Grade/Year	N	P(X=0)	P(X=1)	P(X=2)
469-70	2485	0.826	0.099	0.076
769-70	942	0.860	0.084	0.056
470-71	2469	0.830	0.097	0.073
770-71	876	0.858	0.079	0.063

For simplicity in calculating the theoretical distributions, we assumed each pair had P(X = 0) = 0.84, P(X = 1) = 0.09, and P(X = 2) = 0.07.

Table 6

RESULTS OF CHI-SQUARE TESTS OF DIFFERENCES BETWEEN OBSERVED AND EXPECTED DISTRIBUTIONS OF RESIDUALS FOR MICHIGAN SCHOOLS, REVISED REGRESSIONS

Schools Reporting 8 Times			Schools Reporting 6 Times			Schools Reporting 4 Times		
No. >1	Observed	Expected	No. >1	Observed	Expected	No. >1	Observed	Expected
0	65	80	0	96	99	0	1946	1856
1	29	35	1	32	32	1	328	400
2	24	32	2	22	28	2	230	331
3	12	9	3	7	5	3	62	32
4	7		4	6		4	63	13
5	9		5	4	3			
6	5 (31)	5	6	0 (10)				
7	3							
8	7							
$\chi^2 = 133.9$, Degrees of Freedom = 4			$\chi^2 = 22.7$, Degrees of Freedom = 4			$\chi^2 = 266.6$, Degrees of Freedom = 4		

NOTE: All residuals are derived from a fit of achievement to SES, racial composition, and community type. All Chi-square statistics are significant at the 0.005 level.

that the earlier findings were not mere statistical artifacts resulting from undercontrolling.

To verify this interpretation, we examined the characteristics of the top 15 schools of those reporting scores for all eight test/grade/year combinations (N = 161). These are the 15 schools that are "outside" at least six times out of eight. They average 2 standard deviations above the mean per test; the top two average 2.8 standard deviations above the mean. Since the standard errors of the regressions are approximately three points, the top 15 schools average about six points above the mean even after controlling for SES, racial composition, and community type.

Are six points a lot? Recall that the tests are constructed with an interstudent standard deviation of ten points; the 15 top schools thus raise their children about three-fifths of a standard deviation. Put differently, about 9 percent of the schools are unusually effective in the sense that they seem to raise students on average from the fiftieth to the seventy-second percentile, holding SES, racial composition, and community type constant. In grade equivalent terms, three-fifths of a standard deviation on the fourth-grade Iowa

reading test is about four-fifths of a year. On the seventh-grade test, it is equivalent to about one full grade.*

The differences between these exceptional schools and schools that are below average are, of course, even larger.

Do the top 15 schools have any features in common? Table 7 characterizes the top 15 along the three control variables. Even after controlling for SES, racial composition, and community type, the top 15 schools are predominantly below the state averages on both SES and minority enrollment and tend to be located in rural communities. Only one was above the state SES average; only four had any nonwhite students, and the highest percentage of minority students (9.1) was still below the state average. Only four schools were not located in community type 5, but even these four were in the predominantly rural Michigan peninsula region. How could these results occur after controlling for all three of these variables in the regressions?

There are a number of possible explanations. Suppose the rural schools are simply not as large as the others. If many fewer students were tested in such schools, we would expect the variation of rural school means to be higher. If so, both the underachieving and overachieving schools might be rural, merely because of their smaller sample size.†

To test this explanation, we ran a new regression that controlled for number of students tested as well as SES, racial composition, and community type. The variable for number tested was barely significant,

---

*The grade equivalents are slightly lower for the Iowa mathematics tests: about three-fifths of a year for the fourth-grade test and nine-tenths of a year for the seventh-grade test (Lindquist and Hieronymus 1964). Three-fifths of an interstudent standard deviation on most IQ tests is nine points.

†Suppose all children had an expected score of 50 and a standard deviation of 10 and were randomly distributed throughout the state. The large schools might average 100 children tested; the small schools only 10. The standard deviation of the estimate of a school's true score is a function of the number tested: std. dev. $(\bar{x}) = \sigma/\sqrt{n}$. Thus, the standard deviation of $(\bar{x})$ for schools of size 100 would be $10/\sqrt{100} = 1$; whereas for schools with 10 students tested it would be $10/\sqrt{10}$--more than three times as large. One would therefore expect the smaller schools to have more outliers than larger schools, even though by hypothesis they have "identical" students.

Table 7

CHARACTERISTICS OF 15 OVERACHIEVING MICHIGAN SCHOOLS

School	Mean SES	Community Type	Percent Minority
1	44.6	Rural	0
2	44.1	Rural	0
3	44.6	City	4.9
4	45.9	Rural	0
5	53.2	Rural	0
6	47.1	City	0
7	48.1	Rural	0
8	47.8	Rural	9.1
9	46.1	City	0
10	43.1	Rural	1.8
11	44.2	Rural	0
12	47.1	City	0
13	42.4	Rural	0
14	41.7	Rural	4.3
15	46.3	Rural	0
State averages	49.9	Metropolitan core = 808 City = 331 Town = 659 Urban fringe = 1257 Rural = 1062	11.0

NOTE: Top 15 schools computed from among 161 schools that reported scores for all eight test/grade/year combinations. Regressors were mean school SES, percent minority enrollment, and community-type dummy variables.

and added only 0.005 to the percentage of variation explained (its regression coefficient was -0.008). Similar results occurred when the natural logarithm of the number tested was used in the multiple regression.

We also plotted achievement scores against the number tested for a random sample of 400 schools for the fourth-grade reading data from 1970-1971. The result appears in Fig. 11. The variation of school means is nearly constant to about 100 tested, then it appears to drop slightly. But it is really the average school score that diminishes, not the variability of the data. The schools with larger numbers of students tested also turn out to be those with lower SES, so they also have lower mean achievement scores. Comparing the vertical spread of

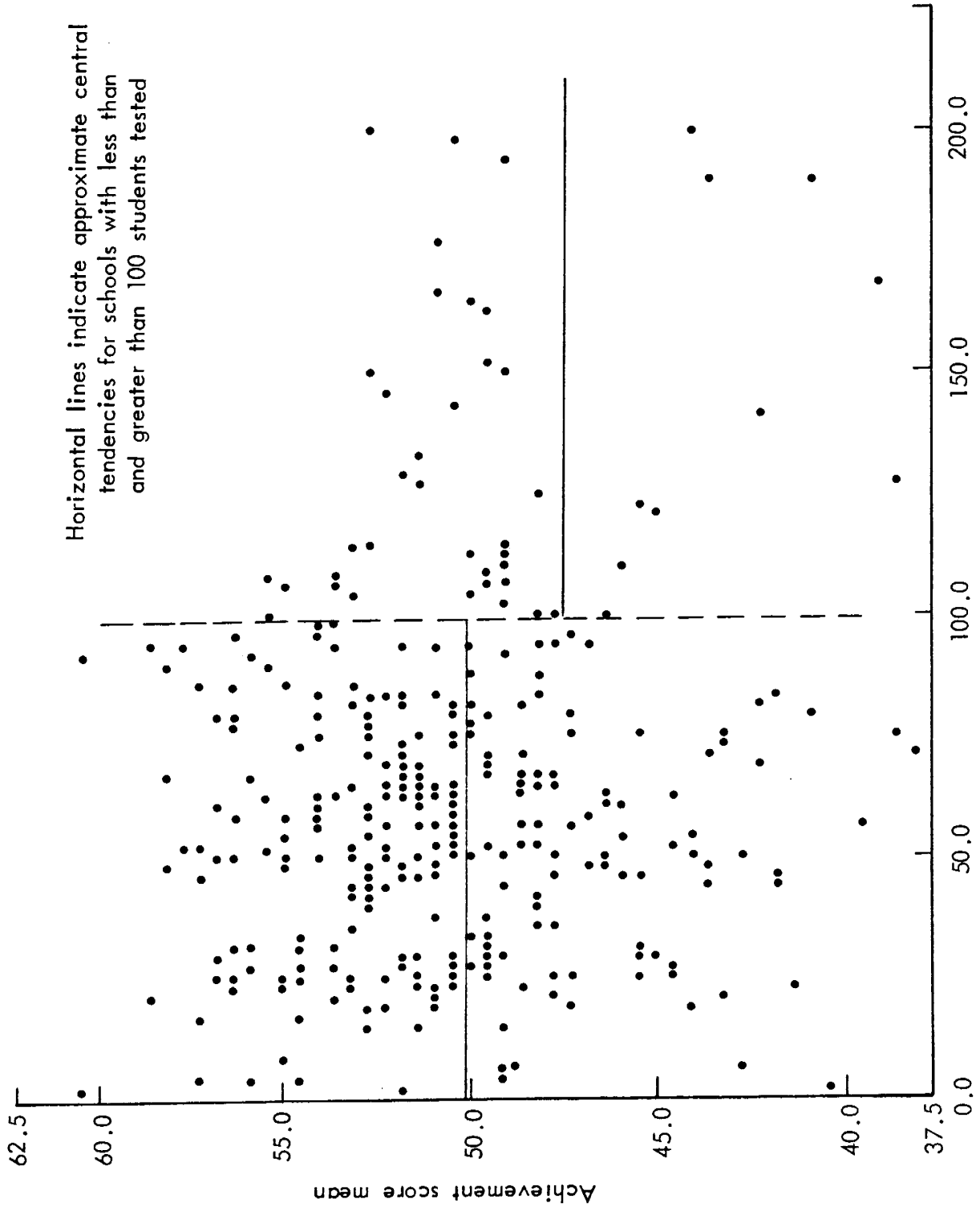


Fig. 11—Plot of achievement scores and number of students tested, random sample of 400 Michigan schools (R470-71)

the data around two hypothetical group averages, one sees that the number of children tested does not seem to correlate with differences in the variability of school means. Most important, the highest- and lowest-achieving schools do not have the smallest number tested. Thus, some variable besides sample size must be adduced to explain the results of Table 7.

Heteroskedasticity Problems

Perhaps some aspect of the fitting procedure accounts for these results. One might expect an equal number of outstanding schools with both low and high SES, with low and high minority enrollment, and from all five community types--but only if achievement has a linear relationship with these independent variables and the residuals have a constant variance. Thus, two characteristics of the data might have caused the exceptional schools to be mostly low SES, white, and rural. First, the dependence of achievement on these variables might not be linear. Second, the residuals from a linear fit might not display a constant variance for all levels of SES, percent minority, and community types.

To find out, we checked the plotted residuals for evidence of curvilinearity and heteroskedasticity. Figure 12 displays the plot of achievement scores against percent minority enrollment. It exhibits curvilinearity--note the U-shape of the plot--but, more important, it shows unequal residual variance as well. The variability of school means among schools with low percentages of minority students is much larger than among those with greater percentages. The set of schools with very low minority enrollment therefore contains both overachieving and underachieving schools to a greater extent than the set with more minority students.

Heteroskedasticity also occurs with respect to community types. For each of the five types, we regressed fourth-grade reading scores for 1970-1971 against SES and racial composition. The results appear in Table 8. The fit is best for the metropolitan core schools ($R^2 = 0.74$); it is by far the worst for the rural schools ($R^2 = 0.07$). Such a fit means, of course, that rural schools exhibit far more variability than other schools, which could account for their disproportional presence among the top 15 schools.

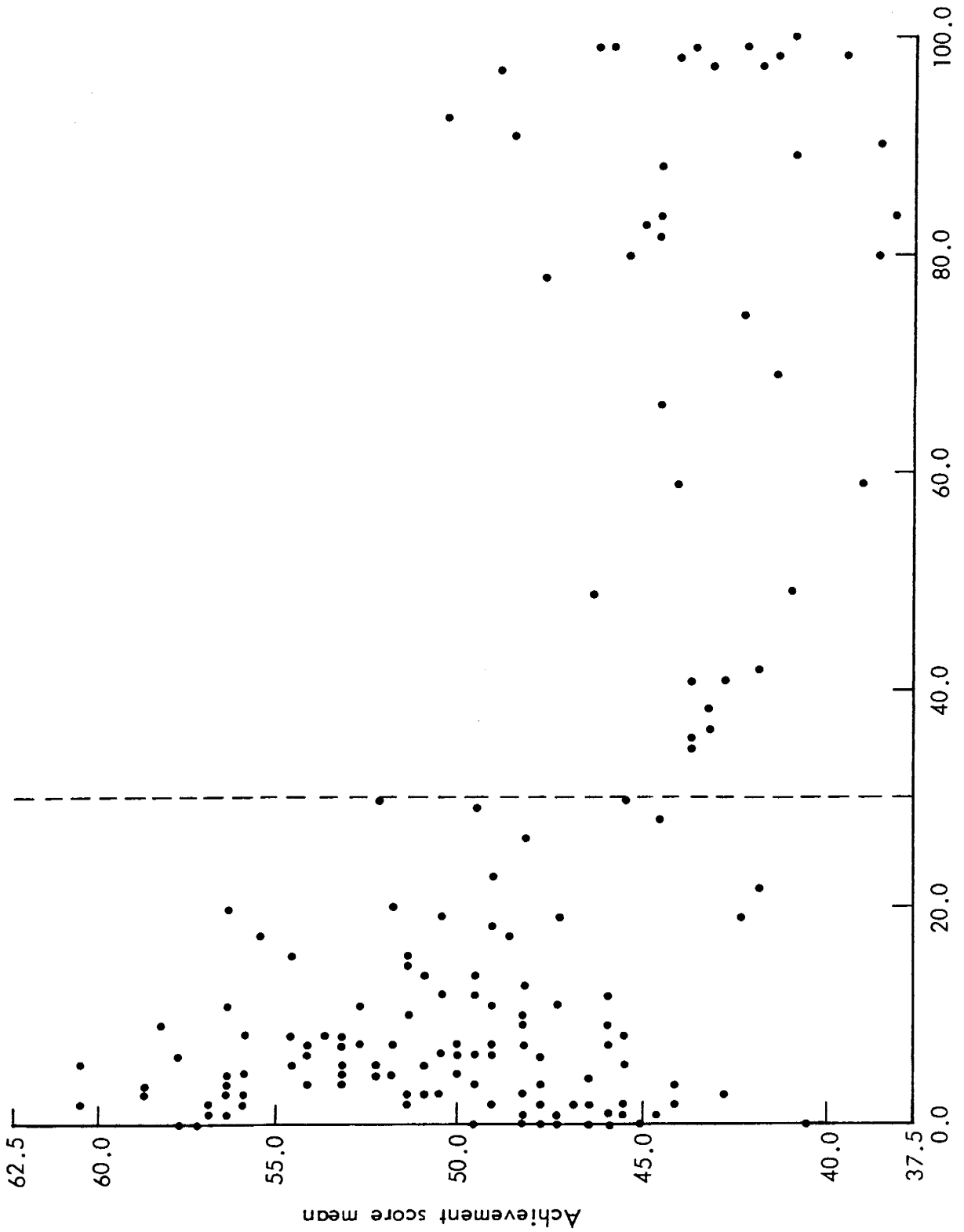


Fig. 12 — Plot of achievement scores and percent minority, random sample of 400 Michigan schools (R470-71)

Table 8

MICHIGAN SCHOOL REGRESSIONS STRATIFIED BY COMMUNITY TYPE, FOURTH-GRADE READING TEST, 1970-1971

Community Type	Equation	R^2	Mean R470	Mean SES	Mean % Minority	Standard Error	Number of Schools
1. Metropolitan core	R470 = 24.01 + 0.54(SES) - 0.06(% minority)	0.74	47.5	47.5	34.3	2.56	474
2. City	R470 = 28.63 + 0.47(SES) - 0.06(% minority)	0.57	51.3	49.9	10.6	2.59	224
3. Town	R470 = 31.70 + 0.40(SES) - 0.09(% minority)	0.27	51.6	50.0	3.4	2.44	374
4. Urban fringe	R470 = 27.55 + 0.46(SES) - 0.05(% minority)	0.51	51.5	51.9	3.4	2.24	824
5. Rural	R470 = 37.45 + 0.28(SES) - 0.02(% minority)	0.07	50.9	49.1	3.2	3.23	547

The results of Table 8, then, reflect heteroskedasticity. The greater variance of rural schools may reflect some important characteristics unique to those schools, but the danger is that it instead reflects imperfections in the SES measures. It is notoriously difficult to compare the economic well-being of rural and urban communities. If the heteroskedasticity is a result of imperfect SES controls, our efforts to locate unusually effective schools may succumb to all manner of misidentification.

Therefore, we tried new regressions to avoid the heteroskedasticity problem. Instead of treating minority enrollment as a continuous variable, we dichotomized it: all schools with less than one-third minority students were assigned a dummy value of one, all others a zero.* All rural schools were excluded. SES and community type for the remaining schools were controlled as before, except that the more valid 1970-1971 SES rating was used for the 1969-1970 data as well.† The results of these regressions appear in Table 9.

---

*We experimented with various nonlinear transformations of the minority enrollment variable on a sample of the data, but they neither significantly raised the percentage of variation explained nor solved the heteroskedasticity problem. Looking again at Fig. 12, one might imagine two sets of data divided by the dotted line. Both sets appear to be randomly distributed around two lines with zero slopes but different intercepts. The data provide a classic case where the use of a dummy variable is warranted: identical slopes but different intercepts.

†The 1970-1971 SES data refer, of course, to the particular fourth- and seventh-graders who provided them. Thus, they do not correspond to the exact data that would have been obtained from the 1969-1970 students. However, we judged that the correspondence would be close enough to warrant the use of the vastly superior 1970-1971 ratings.

Table 9

MICHIGAN SCHOOL REGRESSIONS, OMITTING RURAL SCHOOLS

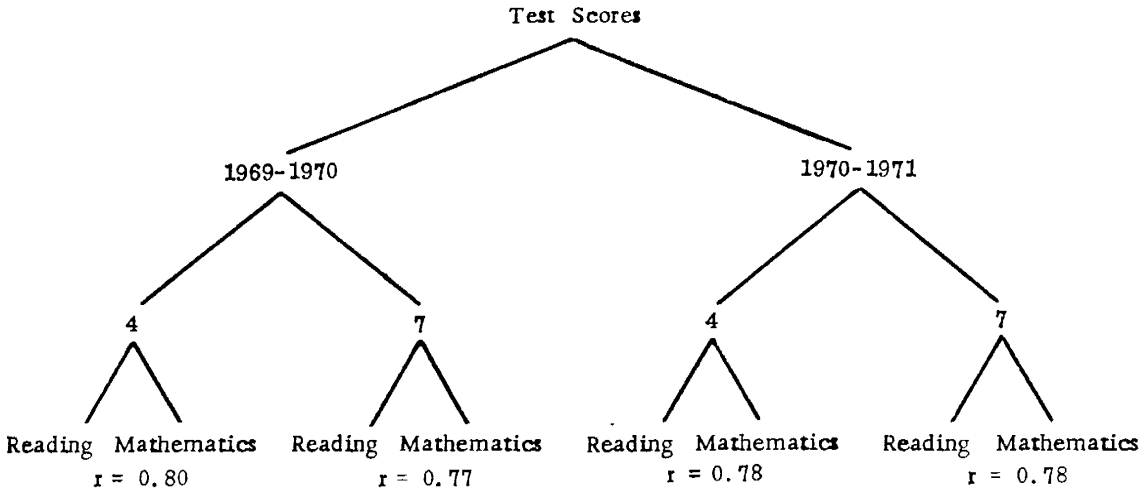
Test	Equation	R^2	Standard Error	Number of Schools
R469-70	Y = 22.18 + 4.12(MIN) + 0.50(SES 4) - 0.78(C1) + 0.27(C2) - 0.78(C4)	0.62	2.56	1836
	325.1 *1200.0* *16.1* *1.6* *21.9*			
M469-70	Y = 22.32 + 4.14(MIN) + 0.50(SES 4) - 0.53(C1) + 0.02(C2) - 0.65(C4)	0.59	2.68	1836
	299.1 *1077.9* *6.7* *0.0* *13.9*			
R769-70	Y = 21.00 + 2.73(MIN) + 0.54(SES 7) + 0.02(C1) + 0.46(C2) - 0.44(C4)	0.75	1.72	480
	86.7 *647.3* *0.0* *2.2* *4.5*			
M769-70	Y = 20.40 + 3.61(MIN) + 0.54(SES 7) - 0.56(C1) + 0.71(C2) - 0.44(C4)	0.72	2.13	480
	99.1 *422.2* *3.4* *3.5* *2.9*			
R470-71	Y = 22.65 + 4.13(MIN) + 0.50(SES 4) - 1.45(C1) + 0.06(C2) - 1.01(C4)	0.66	2.44	1891
	373.6 *1346.8* *62.9* *0.1* *42.3*			
M470-71	Y = 20.33 + 4.39(MIN) + 0.54(SES 4) - 1.26(C1) + 0.16(C2) - 0.79(C4)	0.66	2.55	1891
	385.1 *1433.5* *43.2* *0.5* *23.9*			
R770-71	Y = 20.88 + 3.89(MIN) + 0.53(SES 7) - 0.40(C1) + 0.23(C2) - 0.44(C4)	0.78	1.78	530
	184.5 *661.5* *2.6* *0.6* *4.6*			
M770-71	Y = 20.80 + 4.65(MIN) + 0.53(SES 7) - 1.84(C1) + 0.11(C2) - 0.79(C4)	0.78	2.04	530
	200.8 *491.7* *42.9* *0.1* *11.1*			

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 < 33.3, 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.

A few details about these regressions are noteworthy, although not central to the task of discovering whether consistent outliers still exist. First, they have greater explanatory power. Second, the SES variable has a relatively stable coefficient in all the equations, hovering around 0.5. Third, the new regressions explain seventh-grade scores much better than fourth-grade scores. The R^2s for the former range from 0.72 to 0.78, whereas those for the latter vary between 0.59 and 0.66. Perhaps this result is due to the larger number of students tested in the average seventh grade (about 220) compared to that in the average fourth grade (about 65). The larger the sample size, the smaller the sampling error--and, consequently, the larger the percentage of variation explained by the regression line.

Our main concern, however, is the presence of consistent over-achievers. Table 10 gives the results of the new Chi-square analysis.*

*The residuals were correlated as follows:



The empirical probabilities for each pair are:

Grade/Year	N	P(X=0)	P(X=1)	P(X=2)
469-70	1836	0.808	0.104	0.088
769-70	480	0.831	0.092	0.077
470-71	1891	0.806	0.112	0.082
770-71	530	0.832	0.083	0.085

For simplicity in calculating the theoretical distributions, we assumed each pair had P(X = 0) = 0.82, P(X = 1) = 0.10, and P(X = 2) = 0.08.

Table 10

RESULTS OF CHI-SQUARE TESTS OF DIFFERENCES BETWEEN OBSERVED AND EXPECTED DISTRIBUTIONS OF RESIDUALS FOR 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	4			
7	2				
8	3				
	14				
$\chi^2 = 32.6$, Degrees of Freedom = 4			$\chi^2 = 387.2$, Degrees of Freedom = 4		

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

The results are encouraging. Of the 87 schools reporting eight scores, five have seven or eight scores above one standard deviation, when less than one was expected by chance alone. These schools average 2.0 inter-school standard deviations above the mean on each test. Since the standard errors of the regressions average 2.24, these five schools score four to five points higher than the average school, even after controlling for SES, percent minority enrollment, and community type. Fourteen schools are outside at least four times, compared to only four predicted by the null distribution.

Among schools reporting four scores, 72 are outside each time, compared to 13 predicted by chance. These 72 schools average 1.65 standard deviations above the mean each time, or about 3.7 points higher than the average school's test score. Restating these results in more intuitive terms:

- Of the schools reporting eight scores, about 6 percent move their students from the fiftieth to the sixty-seventh percentile, after allowing for chance variation and controlling for SES, percent minority enrollment, and community type.
- Of the schools reporting four scores, about 2.5 percent move their students from the fiftieth to the sixty-fifth percentile, after allowing for chance variation and controlling for SES, percent minority enrollment, and community type.

We can also make a strong negative statement. No schools in this sample of Michigan schools raise children's scores more than two-thirds of an interstudent standard deviation above the mean.* Thus, after giving schools credit for all residual variation and all possible errors resulting from undercontrolling, no school seems able to raise children, on average, more than about six or seven test points (on tests with mean 50, standard deviation 10).

Characteristics of Exceptional Schools

Our next task was a preliminary investigation of differences between the overachieving schools and the average ones. We did not attempt to locate causal relationships, although of course the existence of regular differences would suggest such relationships. The primary purpose was to see if the schools that seemed to be unusually effective represented truly nonrandom phenomena or merely statistical artifacts.

For both the top 72 schools† and all the Michigan nonrural schools, we tabulated the means and standard deviations for 11 variables, which fell into three groups:

---

*The highest average for any school is 2.96 interschool standard deviations.

†We used the 72 schools that reported four out of four scores at least one standard deviation above the mean. These schools composed about 2.5 percent of the total number of nonrural schools that reported four scores. See Table 10.

1. *School variables.* These included percentage of teachers with more than five years' teaching experience, master's degrees, and salaries exceeding $11,000 annually, as well as the school's pupil/teacher ratio. We hypothesized that, if several of these variables showed significant differences between top schools and the state average, it would provide strong corroborating evidence that these top schools were not mere random deviations.
2. *Background variables.* We examined differences in racial composition and SES for both fourth and seventh grades. Differences along these variables might indicate that the fit was not linear or that heteroskedasticity still existed. Differences here might therefore imply that the outliers were a product of imperfect statistical controls, especially if no school variable differences existed.
3. *Number of students tested.* Although earlier regressions showed that adding the number of students tested as an independent variable did not appreciably help the fit, this variable might still have a significant effect on the outliers. As indicated earlier (pp. 40, 47), we would expect schools with fewer students tested to exhibit greater variation in their mean scores. If the 72 outstanding schools all had very low numbers of students tested, they might rightly be called statistical artifacts instead of unusually effective schools.

In addition, we compared the top and average schools along geographical (both community type and region) and other variables. The results appear in Tables 11 and 12.

Clear and important differences appear in the school-related variables between the top and average schools. Three of the four differences are statistically significant. The unusually effective nonrural schools tend to have lower pupil/teacher ratios, more experienced teachers, and more teachers earning annual salaries over $11,000. They also tend to have more teachers holding master's degrees, but the difference is not significant.

Table 11

COMPARISONS BETWEEN TOP 72 AND ALL NONRURAL SCHOOLS, MICHIGAN DATA

Variables	Top 72 Schools[a]			All Nonrural Michigan Schools			z-statistic[b]
	Mean	Std. Dev.	Number Tested	Mean	Std. Dev.	Number Tested	
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 enrollment	12.9	21.1	72	17.0	31.3	3056	*-1.61*
SES470	49.4	4.4	57	50.1	4.9	1923	*-1.18*
SES770	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 are 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.

Table 12

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

Community Type and Location	Top 72 Schools[a]				All Nonrural Michigan Schools				χ^2[b]
	1	2	3	4	1	2	3	4	
Community type									
Number	21	9	19	24	808	331	659	1257	2.04
Percentage	29.1	11.1	26.3	33.3	26.4	10.8	21.6	41.1	
Region									
Number	21	33	5	13	1271	1413	116	121	40.11*
Percentage	29.1	45.8	6.9	18.0	41.6	46.2	3.8	4.0	

[a]The "top 72" schools are 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 0.005 level.

None of the background variables shows significant differences. The percentage of minority enrollment is close, with $z = 1.61$ ($z_{0.05} = 1.645$). The average nonrural school has 17 percent minority students, whereas the 72 top schools average only about 13 percent. Interestingly, the unusually effective schools are slightly below average on SES, although not significantly so. Given our control variables, the richer schools do not tend to excell. Both the percent minority and SES results are reassuring: errors in the control variables probably did not account for the 72 schools at the top.

The differences between the top and average schools in number of students tested show up as highly significant for fourth grades, but not at all for seventh grades (in fact, in one case the sign is the opposite). The top 72 schools tend to have fewer fourth-graders tested than the average school. However, this difference is only about 10 to 13 students; both sets of schools have enough children tested (50 to 60) to make differences in variability due to sample size extremely small. Thus, the variation in outcomes between the top 72 and average schools does not seem attributable to differences in the numbers of students they had tested.

The results for community type are also encouraging. The outstanding schools tend to be distributed among the four nonrural types randomly, so the dummies for community type are not responsible for the 72 outliers. The regional results do show significant differences between the top and average schools: a disproportionate number of the top schools are located in the northern peninsula region (18 percent compared to 4 percent). Although rural schools are specifically excluded from this analysis, the northern peninsula of Michigan has a decidedly rural flavor, even in its towns and suburbs. The relatively large number of top schools from this region may be another sign that there is a rural component contributing to unusual educational effectiveness. Whether this factor reflects a statistical property of the SES measures, or a real difference in the ability or interest of rural schools to raise achievement scores, is an important question for further research.

Despite the disproportionate number of northern peninsula schools among the top 72, in absolute terms they still represent only a small minority of the overachievers. Most of the 72 are located in southern and central Michigan, and most are in central city or suburban areas. Unusually effective schools--or at least these candidates for that title--are not wholly or even largely a rural or regional phenomenon.

Tables 11 and 12 are extremely important. On balance, they indicate that the outliers we located were indeed nonrandom, and the outstanding schools seem to share certain policy-related attributes. The consistently superior scores of the top 72 schools apparently represent real differences between them and the average school--differences that deserve detailed study at the local level.

NEW YORK CITY DATA

Data were obtained on over 600 New York City elementary schools, representing about 90 percent of the schools in the city.* Achievement score information covered grades 2 through 6 for 1967-1971, although not every school reported scores for each year. We regressed reading scores (in grade equivalents) on the Metropolitan Achievement Test for

*We gratefully acknowledge the efforts of Henry Acland of Harvard's School of Education, who not only provided us with the data but greatly assisted us in numerous stages of their analysis.

each grade and year against the previous grade's score in the previous year. For example, the fourth-grade score in 1969 was regressed against the same school's score for the third grade in 1968. By controlling for previous years' scores, the impact of just the one year of education could be assessed. In effect, this method holds equal the stable background factors like SES and IQ, as well as some school factors relating to the students' previous education.* The regressions are summarized in Table 13.

Table 13

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.

The histograms of the residuals showed no unexpected patterns in the right tails. The distributions were actually tighter than the

*Such regressions present numerous problems, despite their advantages. See Harris (ed.) (1963), Cronbach and Furby (1970), and Harsh in Haggart, Sumner, and Harsh (1972).

normal distribution. Only about 12 percent of the schools had scores more than one standard deviation above the mean, compared to 16 percent for the normal distribution. The histograms revealed very little skewness, as the residuals were symmetrically massed around the zero mean. In short, the histograms gave no *prima facie* evidence of exceptional performers or different educational technologies.

We looked in detail at two years (cross-sectionally across grades) and two grades (longitudinally over years). Four sets of residuals were analyzed cross-sectionally for both 1968 and 1970 to determine whether some schools were overachieving in every grade. Third- and fifth-grade scores were examined longitudinally to find schools that were outstanding in those grades year after year. Complete independence was assumed in the theoretical distribution, with P(score > 1) = 0.12. Table 14 gives the results of our Chi-square analyses.

The results are suggestive, but not as clear as in the Michigan case. Two of the Chi-square statistics are much lower than any in the Michigan analyses, though still statistically significant. For 1970, for example, it is difficult to argue that there are more outstanding schools than chance alone would allow. Seven schools are outside at least three times, but three would have been expected just from random variation. The significant Chi-square statistic is partly due to differences at the low end and in the middle of the distribution. Similar remarks apply to the fifth-grade results.

However, in both the 1968 and third-grade analyses, twelve schools were above one standard deviation at least three out of four times, but only three were expected in 1968 and only four in the third grade. The Chi-square statistics are healthier, although again their size is not just due to overachievers.

Summarizing the results obtained from the New York City data:

- For 1970 (all grades) and fifth grades (all years), we find no significant evidence that any schools perform consistently better than average, after allowing for random variation.

Table 14

RESULTS OF CHI-SQUARE TESTS OF DIFFERENCES BETWEEN OBSERVED AND EXPECTED DISTRIBUTIONS OF RESIDUALS FOR 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
$\chi^2 = 33.6,^{\dagger}$ Degrees of Freedom = 3			$\chi^2 = 10.4,^{\ddagger}$ 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
$\chi^2 = 7.8,^{*}$ Degrees of Freedom = 3			$\chi^2 = 21.4,^{\dagger}$ 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.

- For 1968 (all grades), about 2.1 percent of the schools seem to raise their children about two-thirds of a grade level above that expected from their 1967 scores, after allowing for random variation.
- For third grades (all years), only about 1.4 percent of the schools seem consistently able to raise their third-graders about half a grade level above that expected from their second-grade scores, after allowing for random variation.
- Only a minority of the overachieving 1968 and third grades are able to raise their children's scores every year. Moreover, the New York City data indicate that no school is consistently able to raise its children an entire grade level over that expected from earlier achievement.

Working with growth scores--the residuals when one year's scores are controlled for scores from the previous grade and year--is tricky. A school's fifth-grade residual, for example, could indicate either how well the fifth grade performed or how badly the previous year's fourth grade did. In the extreme, a school that seems to have an exceptional fifth grade year after year may just have a consistently bad fourth grade. But we found in two sets of four residuals that chance alone would have predicted almost the number of "outliers" that occurred, and that in any case the "outliers" were not far from the mean. Examining the New York City data for unusually effective grades and for schools that performed well on all grades for any one year uncovered at best equivocal evidence.

PROJECT TALENT DATA

The original 1960 Project Talent survey comprises the largest compilation of educational achievement data. We obtained a sample of 858 high schools from this original data set, including 708 that reported scores for both ninth and eleventh grades. The scores for composite achievement tests, similar to the Scholastic Achievement Tests (mean = 500, standard deviation = 100 for twelfth graders), were regressed

against SES indexes for each school. The results of the two regressions appear in Table 15.

Table 15

PROJECT TALENT SCHOOL 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(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(SES)	429.6	0.34	72.1	820	493.1	89.0	96.2	7.1

As in the other data we have examined, the histograms of the residuals are unimodally massed with well-behaved tails. The distributions are tighter than a normal distribution: the actual probability of a school scoring more than one standard deviation above the mean is 0.118 for the ninth grade and 0.134 for the eleventh grade (compared to around 0.159 for the normal distribution). Using these actual probabilities, the null distribution assuming independence was computed and compared with the actual results over both tests. These comparisons appear in Table 16.

Table 16

RESULTS OF CHI-SQUARE TEST OF DIFFERENCE BETWEEN OBSERVED AND EXPECTED DISTRIBUTIONS OF RESIDUALS FOR 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.

No significant difference is evident between the expected and observed distributions, even though SES is the only control variable. Fifteen schools instead of eleven are "outside" in both the ninth and eleventh grades, but this number is within the range allowed by chance alone. Neither this result nor the histograms provide any support for the idea that unusually effective high schools exist.

NEW YORK STATE AND YARDSTICK PROJECT DATA

We investigated two other potential sources of information about unusually effective schools--the New York state Basic Educational Data System (BEDS) and the Yardstick Project in Cleveland, Ohio.

For purposes of analysis, the state of New York generally aggregates data on its approximately 6000 public schools at the district level. However, through the cooperation of the University of the State of New York, we obtained school-level information for a one-tenth random sample of the state's elementary schools. Every year since 1966, New York schools have administered "reading readiness" tests to entering first-graders and achievement tests to third- and sixth-graders. We attempted to follow one "cohort" of children from their 1966 first-grade reading readiness scores through their 1968 third-grade scores and on to their 1971 sixth-grade results. Using the 1966 scores as controls in regressions against the 1968 and 1971 scores, and the 1968 results as controls for the 1971 outcomes in a second regression, we examined the histograms of residuals to see if any schools seemed to account for unusually large reading increases. Unfortunately, data imperfections appeared to exist in our sample: six schools had very extreme values in the third grade compared with the rest of the schools and with their own first- and sixth-grade scores. No clerical or theoretical explanation was apparent, and otherwise the histograms appear quite normal.

But if unexplained imperfections limited the usefulness of our sample of this data, it is clear that in future years BEDS information will be extremely valuable for educational research. Since the 1972 sixth-grade data were not yet collated, and the data system began only in 1966, we could not yet compare first- to sixth-grade growth rates

longitudinally. However, the reading readiness score is an attractive regressor variable: it picks up much of the constant personal and socioeconomic factors that influence achievement, without showing any school effects. Furthermore, by using the sixth-grade level of achievment as the response variable, one can evaluate the total impact of an elementary school on a child, without encountering the vexing problem of interpreting series of year-to-year residuals. Thus, although the New York data were of only marginal use in this study, future efforts along the same lines should find them a fruitful source.

The other achievement information we examined came from the Yardstick Project in Cleveland. Yardstick contracts its data analysis services to about 34 school districts in Ohio and other states. Its clientele varies from year to year, as do the clients' data requests: some ask for analyses of lower elementary grades and some upper, and over different spans of time. The Yardstick data bank stratifies school data in interesting ways. For instance, it provides growth-per-year scores stratified by five IQ levels and five categories of father's occupation. But our investigations revealed that the data were simply too heterogeneous to be useful in comparing the effects of different school policies or locating unusually effective schools.

ALTERNATIVE LEVELS OF AGGREGATION

A few final explorations were made, experimenting with levels of aggregation other than schools. Indeed, one of the real questions raised by this and other studies is, at what level should one look for unusual educational effectiveness?*

Using data from 628 New York state school districts, we searched for evidence of unusually effective *districts*. Third- and sixth-grade achievement scores in reading and mathematics for 1969-1970 and 1970-1971 (a total of eight scores per district) were individually regressed by the University of the State of New York against a host of background variables. Table 17 gives their regression results.

The histograms of the residuals once again reveal no surprises. The residuals of the reading and mathematics scores are correlated in each grade/year combination, as in the Michigan case. Thus, allowance

*See, for example, Hanushek (1970), Kiesling (1969), and Averch, et al. (1972).

Table 17

NEW YORK STATE SCHOOL DISTRICT REGRESSIONS

Test	Equation	R^2
R369	$Y = 16.66 + 0.16(S68) + 0.04(FTV68) - 7.20(PR70) - 0.67(D68) + 0.28(R166,67) - 0.13(R1sd66,67)$	0.35
R370	$Y = 14.76 + 0.13(S68) + 0.03(FTV68) - 5.22(PR70) - 1.61(D68) + 0.30(R167,68) - 0.11(R1sd67,68)$	0.33
A369	$Y = 16.80 + 0.17(S68) + 0.05(FTV68) - 11.42(PR70) - 3.50(D68) + 5.24(D68^2) - 0.11(D68)(FTV68) + 0.27(R166,67) - 0.12(R1sd66,67)$	0.35
A370	$Y = 14.94 + 0.16(S68) + 0.04(FTV68) - 9.79(PR70) - 8.48(D68) + 8.73(D68^2) - 0.12(D68)(FTV68) + 0.30(R167,68) - 0.14(R1sd67,68)$	0.41
R669	$Y = 10.10 + 0.02(FTV68) + 0.70(R366) + 0.31(R3sd66) - 7.82(PR70) - 2.90(G68) + 0.13(D68) - 0.04(D68^2) + 0.08(R166,67)$	0.62
R670	$Y = 6.44 + 0.02(FTV68) + 0.69(R367) + 0.42(R3sd67) - 8.01(PR70) - 2.73(G68) - 1.41(D68) - 0.43(D68^2) + 0.12(R167,68)$	0.63
A669	$Y = 4.49 + 0.02(FTV68) + 0.58(A366) - 0.16(R3sd66) - 0.45(A3sd66) - 6.93(PR70) - 0.65(D68) + 0.10(R166,67)$	0.40
A670	$Y = 0.84 + 0.02(FTV68) + 0.61(A367) + 0.15(R3sd67) + 0.22(A3sd67) - 7.35(PR70) + 1.22(D68) + 0.11(R1sd67,68)$	0.44

NOTE: These data were taken from the University of the State of New York (1972), pp. 17f. R369 stands for third-grade reading scores in 1969. The other test symbols are interpreted similarly. The remaining abbreviations are explained in the matrix of regressor-response variable combinations below:

	Regressor Variables							
	Third Grade				Sixth Grade			
	Reading		Arithmetic		Reading		Arithmetic	
Regressor Variables	1969	1970	1969	1970	1969	1970	1969	1970
First-grade reading readiness								
1966-1967 mean (R166,67)	X		X		X		X	
1966-1967 std. dev. (R1sd66,67)	X		X					
1967-1968 mean (R167,68)		X		X		X		X
1967-1968 std. dev. (R1sd67,68)		X		X				
Third-grade reading								
1966 mean (R366)					X			
1966 std. dev. (R3sd66)					X		X	
1967 mean (R367)						X		
1967 std. dev. (R3sd67)						X		X
Third-grade arithmetic								
1966 mean (A366)							X	
1966 std. dev. (A3sd66)							X	
1967 mean (A367)								X
1967 std. dev. (A3sd67)								X
Other variables								
1968 full tax value (FTV68)	X	X	X	X	X	X	X	X
1970 pupil race (PR70)	X	X	X	X	X	X	X	X
1968 enrollment size (S68)	X	X	X	X				
1968 enrollment growth (G68)					X	X		
1968 enrollment density (D68)	X	X	X	X	X	X	X	X
Density squared ($D68^2$)			X	X	X	X		
Interaction term (D68) × (FTV68)			X	X				

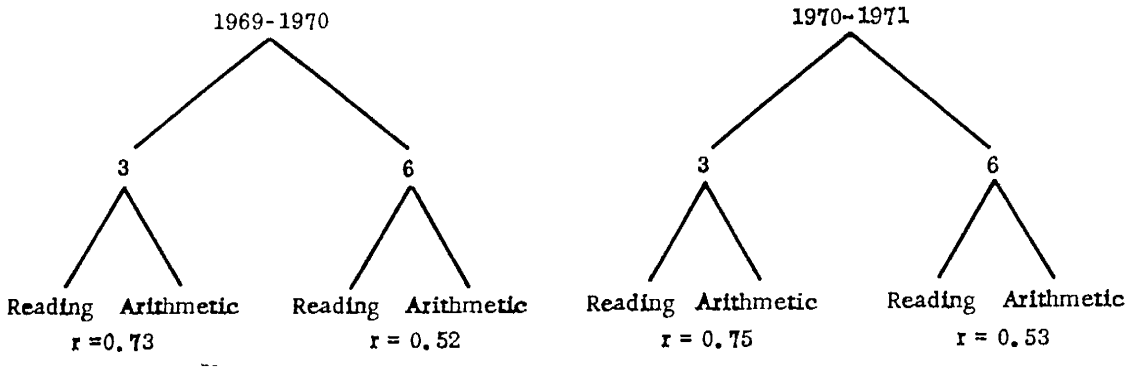
once again must be made for the correlations among these pairs when computing the theoretical distribution.*

The Chi-square test shows strong evidence that some districts are persistent outliers. Table 18 presents the results. Fifteen districts report scores at least one standard deviation above the mean six out of eight times or more; not even one would have been predicted by chance. Thirty districts are "outside" five out of eight times, compared to only four predicted by chance. Unfortunately, the University of the State of New York does not give the size of the standard errors of the regressions, so it is impossible to estimate how much difference, in terms of achievement scores, the unusual districts make. However, these districts are promising candidates for further detailed investigation.

We also wondered if some *grades* (instead of schools) might be unusually effective. Our analysis of the New York City schools included tests to see whether unusually effective third and fifth grades existed (pp. 55-57); we found some evidence of exceptional third grades, but none of outstanding fifth grades. Overall, the exploration was not exceptionally encouraging.

---

*The tests were correlated as follows:



We calculated the theoretical distribution using the following empirical probabilities:

Grade/Year	N	P(X=0)	P(X=1)	P(X=2)
369-70	628	0.787	0.146	0.067
669-70	628	0.807	0.118	0.075
370-71	628	0.788	0.161	0.051
670-71	627	0.788	0.171	0.041

Table 18

RESULTS OF CHI-SQUARE TEST OF DIFFERENCE BETWEEN OBSERVED AND EXPECTED DISTRIBUTIONS OF RESIDUALS FOR NEW YORK STATE SCHOOL DISTRICTS

Grades 3 and 6, Reading and Arithmetic, 1969-1970 and 1970-1971

No. >1	Observed	Expected
0	318	247
1	119	186
2	81	125
3	44	48
4	35	16
5	15	
6	8	4
7	5	
8	2	
	30	

χ^2 = 233.4, Degrees of Freedom = 5

NOTE: The Chi-square statistic is significant at the 0.005 level.

In addition, we reanalyzed the Michigan results for evidence of unusually effective grades. For the schools that reported eight scores (including rural schools), we searched for evidence of exceptional fourth or seventh grades apart from the combined effect school policies might have on both grades. We broke down the residuals into four groups of four:

- Fourth grade (both years, reading and mathematics).
- Seventh grade (both years, reading and mathematics).
- 1969-1970 (fourth and seventh grades, reading and mathematics).
- 1970-1971 (fourth and seventh grades, reading and mathematics).

If the two "grade groups" are marked by substantially more persistent outliers than are the two "year groups," then influences at the grade level (rather than the school level) might be accounting for unusual effectiveness.

Once again we used Chi-square analysis, and the results appear in Table 19. There do seem to be unusually effective grades, but no more than unusually effective years--i.e., schools that, over both grades, obtain unusually good results for a single year.*

Table 19

COMPARING GRADE EFFECTS AND SCHOOL EFFECTS FOR THE MICHIGAN DATA

Grade Effect for 4th Grade, 1969-1971			Grade Effect for 7th Grade, 1969-1971		
No. >1	Observed	Expected	No. >1	Observed	Expected
0	1615	1572	0	605	571
1	331	371	1	72	108
2	215	304	2	56	84
3	64	32	3	25	7
4	66	14	4	16	3
$\chi^2 = 262.4$, Degrees of Freedom = 4			$\chi^2 = 124.2$, Degrees of Freedom = 4		
School Effect for 4th and 7th Grades, 1969-1970			School Effect for 4th and 7th Grades, 1970-1971		
No. >1	Observed	Expected	No. >1	Observed	Expected
0	210	241	0	99	137
1	52	52	1	39	29
2	47	40	2	26	24
3	17	4	3	11 } 28	3
4	14	1	4	17 }	
$\chi^2 = 163.5$, Degrees of Freedom = 4			$\chi^2 = 201.9$, Degrees of Freedom = 3		

NOTE: The residuals were obtained from the regressions in Table 4. All Chi-square statistics are significant at the 0.005 level.

*The empirical probabilities from which the theoretical distributions were computed are given in the note on p. 38. Note that the seventh grade has fewer schools outside one or both times than the fourth grade; thus, the theoretical distributions for the grade effects were computed using the empirical grade probabilities and not by using the probabilities of an average pair of scores.

Therefore, school policies are probably accounting for the persistent outliers in the fourth and seventh grades considered separately. Table 19 shows no distinct evidence that studying unusually effective grades would be worthwhile.

However, many other levels of aggregation could be explored, such as the individual classroom or teacher. Examining particular programs might also be fruitful, since programs cut across grades and schools but are not necessarily applied uniformly throughout them. Our study concentrated, of course, on schools, but the search for unusual educational success should perhaps take broader (e.g., district) and narrower (e.g., classroom) perspectives as well.

In addition to the level of aggregation, there also exists the important question of the appropriate measures of school effects. Like most who have studied school effectiveness, we have used schools' *mean* achievement-test scores. Average results, however, may not indicate very accurately how well a school advances educational policy objectives.

For example, a major educational objective today is to provide equality of opportunity to children and thus facilitate social mobility. A school may be unusually successful in achieving these goals because it provides some children with chances to succeed far beyond what would be expected from their socioeconomic backgrounds. However, the success of such students may be swamped in the statistics by the greater number who make no special gains.

Another educational goal requires that schools place special emphasis on imparting some minimum levels of learning to all students, or, alternatively, that society place special value on schools that can successfully remedy educational handicaps. This policy focus does not regard the achievements of all students as equivalent, but places a higher value on the gains of students with disadvantages--i.e., those on the lower end of the achievement-score distribution. Again the school mean will not be a helpful measure, since algebraically it treats a gain of five points the same no matter who attains it--whether one moves from 30 to 35 or from 65 to 70.

Other dimensions of educational policy are also imperfectly reflected in the use of average scores. For example, we value programs

that enable academically gifted children to maximize their potentials. Mean test scores will hardly reflect an atypical number of very high achievers in schools with mostly average students.

The important point for research is that, considering the objectives of public policy toward education, researchers should examine aspects of the intraschool distributions of test scores other than the mean. Three potentially useful statistics are the variance, the skewness, and the proportion of students above some minimum level of achievement. The variance of residuals indicates how far students deviated from expected scores and thus may be a useful proxy for educational mobility. The skewness statistic gives greater weight to gains on the tails than to gains around the mean and may therefore reflect success with low and high achievers.* The proportion of students above some minimum is an indicator of school success in providing basic skills. Including some such measures related to intraschool distributions raises challenging analytical and statistical problems, but our efforts to date indicate that these are surmountable.† Considering the lack of congruence between the objectives of educational policy and the use of average test scores, it appears highly desirable to include analysis of the spread and shape of school achievement results in future research.

---

*The more positively skewed the distribution, *ceterus paribus*, the better the school is doing with the high and low achievers at the expense of students around the mean. It is not important that the particular statistics used be the variance and the skewness, but that some measure of the spread and shape of the intraschool distributions be considered along with the mean.

†The authors are currently investigating these new uses for achievement scores in educational evaluation.

IV. CONCLUSIONS

This report comprises a series of exploratory investigations of the existence of unusually effective schools. We assumed that school effectiveness could be gauged by mean performance on standardized achievement tests. Because nonschool factors significantly affect achievement scores, we did not define "effectiveness" in absolute terms: rather, we examined school performance after controlling for a variety of background variables. We assumed a simple model of the form

$$\text{Achievement score} = f(\text{background, school}).$$

Then, using multiple regression techniques, we experimented with various controls for the background influences, using such measures as previous test scores, SES, minority enrollment, community type, and others, in varying combinations.

Instead of concentrating on the properties of the regression lines obtained, we studied the *residuals* from those lines. Although we recognized that residual variation would stem from sources other than different school effects, we decided on the basis of previous studies and the nature of our problem to proceed as if all variation after the fit resulted from the schools (and some random factors). Then, by examining series of residuals from the same schools both cross-sectionally and longitudinally, we tried to eliminate random effects and locate true outliers that would be candidates for further, more detailed study. Although the existence of such outliers would not necessarily be due to unusually effective schools, to discover no outliers under such circumstances would be a very strong negative result. Finally, where possible, we investigated the apparent overachievers for shared characteristics, in a further attempt to separate random from nonrandom outliers.

Jencks and others have shown how tight the distribution of school achievement scores is after controlling for nonschool background factors

(Jencks, et al. 1972). Our various regressions support that finding. Even in regressions where the percentage of variation explained is not high, the standard errors are small compared with the interstudent standard deviation of the achievement test. Figure 13 contrasts the interstudent distribution of test scores with a typical interschool distribution controlled for background factors. There is more variation within schools than between them. This general result is true whether our primary control is previous test scores or an explicit socioeconomic measure.

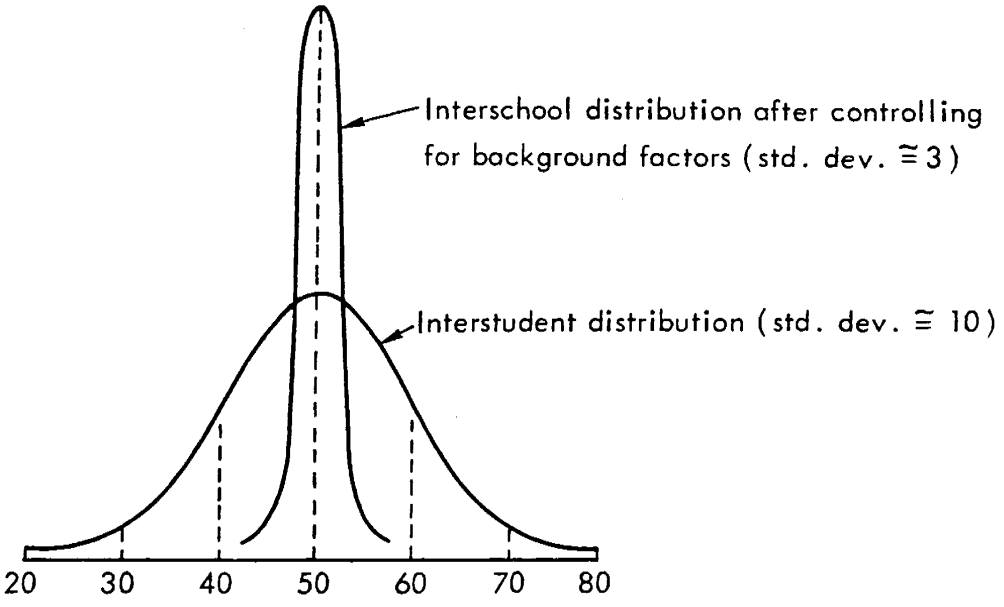


Fig. 13 — Schematic representation of interschool and interstudent variation of achievement scores

The second unanimous finding from all our disparate explorations involves the normality of the distribution of residuals. The overwhelming impression is one of smooth, well-behaved, random variation. There is no evidence of extreme outliers nor of discontinuity in educational effectiveness among schools.

The results of analyzing school performance over many grades and years are more suggestive, but quite mixed. They can be summarized as follows:

The Michigan data provide evidence of unusually effective schools.

1. Including rural schools, the Chi-square tests showed a greater number of consistently overachieving schools than chance alone would allow. For example, among the 161 schools that reported scores for all eight test/grade/year combinations, 15 were at least one standard deviation above the mean six out of eight times (less than one was expected by chance). That is, about 9 percent of the schools seemed able to raise their students, on average, by an amount equal to an increase from the fiftieth to the seventy-second percentile, given equal background factors. However, most of these outstanding schools were rural and all white, even after controlling for community type and percent minority, which reflects heteroskedasticity in the control variables. By running regressions stratified by community type, we found that our regressor variables could explain only 7 percent of the variation among rural schools, compared with 50 to 60 percent for the other four community types combined. These results may reflect unique characteristics of rural schools, or they may result from imperfect measures for SES.
2. Not including the rural schools, we found evidence of consistent overachievers in two of the four Chi-square tests. For example, among the 2131 schools that reported scores for four test/grade/year combinations, 72 were at least one standard deviation above the mean all four times (13 were expected by chance). In other words, about 2.5 percent of these schools seemed able to improve their students an amount equivalent to an increase from the fiftieth to the sixty-fifth percentile, given equal background factors.

   Furthermore, these 72 schools differed significantly from the average nonrural school on three out of four school-related factors. 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 annually. 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. Neither could differences in nonschool factors, although 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 grouping may be evidence of some regional/rural factor contributing to unusual effectiveness.

For the New York City data, the results were equivocal. Although in one year for all grades and for one grade over time, we found some evidence of consistent overachievers, random variation could account for almost all the outliers observed in the other year and grade. Furthermore, the consistent overachievers that were identified averaged only 1.5 interschool standard deviations above the mean, not so much as in the Michigan schools. Very few schools indeed were above one standard deviation every time.

The Project Talent data showed no evidence of consistently overachieving schools apart from that predicted by chance alone. 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 briefly examined two other levels of aggregation and found evidence strongly suggesting outstanding *districts*. Among the 627 districts we studied, thirty were above one standard deviation at least five out of eight times, while fewer than four such districts were expected by chance. (Fifteen were outside six or more times, compared with one district predicted by the null hypothesis.) Unfortunately, the state of New York did not provide information that would allow us to gauge how far these districts were able to raise their students' scores in percentiles or as compared with other students.

We also looked for unusually effective *grades*, but found little evidence to encourage further investigation of this hypothesis. Although there were sometimes more outstanding grades than chance would predict, the amount of deviation was consistent with the notion that school effects rather than grade effects accounted for these outliers.

When we did identify a group of overachieving schools, they comprised from 2 to 9 percent of the sample and averaged about 0.4 to 0.6 interstudent standard deviations above the mean per test. Since our analyses of different data sets involved different grades, different achievement tests, and different control variables, positive generalization about the number and magnitude of exceptional schools is impossible. However, our exploratory use of so many different kinds of variables makes a *negative* statement all the more powerful: however simple the controls, whatever grades, and even positing that all residual variation represented school effectiveness, no school was consistently able to raise its students' achievement scores more than about 0.8 interstudent standard deviations.

Thus, the policy relevance of this exercise comes down to a question of valuation. We located some schools that were statistically "unusual," but whether they can be considered unusually *effective* depends on one's subjective scale of magnitude. Is half an interstudent standard deviation a lot? How much better would a school have to be before it would be worth studying in detail for purposes of replication? And can present educational research and development policy, with its emphasis on diffusing and imitating best practice, base itself on interschool differences of these magnitudes? Such normative questions clearly transcend statistics.

Nonetheless, moving away from average effects in educational research and policymaking does seem worthwhile. We have located schools and districts that consistently perform better than their peers. It is probably worthwhile to continue such research, and to begin looking for unusually effective classrooms and programs. The methodology developed in this report should prove useful in such efforts. As educational researchers continue to devise new measures of school outcomes,

and as they begin to focus 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.

BIBLIOGRAPHY

Acland, Henry D., "School Effects: An Analysis of New York Elementary Schools" (unpublished).

Averch, Harvey A., et al., *How Effective is Schooling?* The Rand Corporation, R-956-PCSF/RD, 1972.

Barro, Stephen M., "An Approach to Developing Accountability Measures for the Public Schools," *Phi Delta Kappan*, Vol. 52, No. 4, December 1970, pp. 196-205.

Bowles, Samuel, and Herbert Gintis, "IQ and the U.S. Class Structure," *Social Policy* (forthcoming).

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, pp. 2-24.

Brown, Byron W., "Achievement, Costs and the Demand for Public Education," *Western Economic Journal*, Vol. 10, No. 2, June 1972, pp. 198-219.

Coleman, James S., et al., *Equality of Educational Opportunity*, U.S. Department of Health, Education, and Welfare, Office of Education, OE-38001 (U.S. Government Printing Office, Washington, D.C., 1966).

Cronbach, Lee J., and Lita Furby, "How We Should Measure 'Change'--Or Should We?" *Psychological Bulletin*, Vol. 74, No. 1, July 1970, pp. 68-80.

Dyer, Henry S., "The Measurement of Educational Opportunity," in Frederick Mosteller and Daniel P. Moynihan (eds.), *On Equality of Educational Opportunity*, Random House (Vintage Books), New York, 1972, pp. 513-527.

Guthrie, James W., "What the Coleman Reanalysis Didn't Tell Us," *Saturday Review*, July 22, 1972.

Hanushek, Eric A., *The Value of Teachers in Teaching*, The Rand Corporation, RM-6362-CC/RC, December 1970.

Harris, C. W. (ed.), *Problems in Measuring Change*, University of Wisconsin Press, Madison, 1963.

Harsh, J. R., "Problems of Measuring Student Gain," in S. A. Haggart, G. C. Sumner, and J. R. Harsh, *A Guide to Performance Contracting: Technical Appendix*, The Rand Corporation, R-955/2-HEW, 1972.

Hilgard, Ernest R., "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*,

92d Cong., 2d sess., January 1972, Appendix 1 to hearings before the Select Subcommittee on Education on H.R. 3606, to create a National Institute of Education, U.S. Government Printing Office, Washington, D.C., pp. 51-57.

Jencks, Christopher S., "The Coleman Report and the Conventional Wisdom," in Frederick Mosteller and Daniel P. Moynihan (eds.), *On Equality of Educational Opportunity*, Random House (Vintage Books), New York, 1972, pp. 69-115.

-----, et al., *Inequality*, Basic Books, New York, 1972.

Kiesling, Herbert J., *The Relationship of School Inputs to Public School Performance in New York State*, The Rand Corporation, P-4211, October 1969.

Lindquist, E. F., and A. N. Hieronymus, *Iowa Tests of Basic Skills: Manual for Administrators, Supervisors, and Counselors*, Houghton Mifflin, Boston, 1964.

Mayeske, George W., et al., *A Study of Our Nation's Schools*, U.S. Department of Health, Education, and Welfare, Office of Education (U.S. Government Printing Office, Washington, D.C., 1969).

Shaycoft, Marion F., *The High School Years: Growth in Cognitive Skills*, Interim Report 3, Project 3051, Contract OE-6-10-065, American Institutes for Research, Palo Alto, California, 1967.

Smith, Marshall S., "*Equality of Educational Opportunity:* The Basic Findings Reconsidered," in Frederick Mosteller and Daniel P. Moynihan (eds.), *On Equality of Educational Opportunity*, Random House (Vintage Books), New York, 1972, pp. 230-342.

Stevens, William K., "Test Expert Calls I.Q. and Grade Equivalency Scores 'Monstrosities'," New York *Times*, March 23, 1971.

Tukey, John W., *Exploratory Data Analysis* (limited preliminary edition, 3 vols.), Addison-Wesley Publishing Company, Reading, Mass., 1970.

University of the State of New York, State Education Department, Bureau of School Programs Evaluation, *New York State Performance Indicators in Education, 1972 Report*, Albany, New York, September 1972.

U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Educational Statistics (Division of Data Analysis and Dissemination), "Characteristics Differentiating Under- and Over-achieving Schools," Washington, D.C., March 1968 (unpublished).

Yarnold, James K., "The Minimum Expectation in χ^2 Goodness of Fit Tests and the Accuracy of Approximations for the Null Distribution," *Journal of the American Statistical Association*, Vol. 65, No. 330, June 1970, pp. 864-886.