DOCUMENT RESUME

ED 117 198 95 TM 005 058

AUTHOR	Klitgaard, Robert E.
TITLE	Looking for the Best: Identifying Exceptional Performers in Education and Elsewhere. TM Report 50.
INSTITUTION	ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
REPORT NO	ERIC-TM-50
PUB DATE	Dec 75
CONTRACT	NIE-C-400-75-0015
NOTE	19p.; For related documents, see ED 085 402 and 409
AVAILABLE FROM	ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, Princeton, N.J. 08540 (free while supplies last)
EDRS PRICE	MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS	Data Analysis; *Demonstration Programs; *Educational Quality; *Identification; Mathematical Models; Multiple Regression Analysis; Program Effectiveness; *Program Evaluation; *Statistical Analysis
IDENTIFIERS	*Statistical Outliers

ABSTRACT

To the data analyst, outliers can present both a problem and an opportunity. Stray or outlying observations can severly distort estimates of a distribution's central tendency (like the mean) and estimates of one variable's relationship to another (like the regression coefficient). These problems are frequent and serious, and as a result, increasing numbers of statisticians are developing new estimating procedures that are robust in the face of outliers. But outliers may also present an opportunity. An unusual observation may indicate the existence of a process not operating in the rest of the distribution. In education, for example, an outlier may turn out to be an unusually effective school, perhaps one worthy of emulation throughout the educational system. Finding the best by locating outliers may be particularly important in the evaluation of public policies. Can one find exceptionally good police forces and study the causes for their success? What about outstanding rural development projects, exceptional hospitals, and unusally effective manpower training programs? Indeed, one may suggest a worthwhile general rule for policy evaluations: For exceptions to the general rules, include a search for unusual performers. (Author)

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

TM REPORT 50 DECEMBER 1975

LOOKING FOR THE BEST: IDENTIFYING EXCEPTIONAL PERFORMERS IN EDUCATION AND ELSEWHERE*

Robert E. Klitgaard

INTRODUCTION

To the data analyst, outliers can present both a problem and an opportunity. Stray or outlying observations can severely distort estimates of a distribution's central tendency (like the mean) and estimates of one variable's relationship to another (like the regression coefficient). These problems are frequent and serious, and as a result, increasing numbers of statisticians are developing new estimating procedures that are robust in the face of outliers (1). But outliers may also present an opportunity. An unusual observation may indicate the existence of a process not operating in the rest of the distribution. In education, for example, an outlier may turn out to be an unusually effective school, perhaps one worthy of emulation throughout the educational system (2). Finding the best by locating outliers may be particularly important in the evaluation of public policies. Can one find exceptionally good police forces and study the causes for their success? What about outstanding rural development projects, exceptional hospitals, and unusually effective manpower training programs? Indeed, one may suggest a worthwhile general rule for policy evaluations: For exceptions to the general rules, include a search for unusual performers.

Why Study Exceptional Performers?

One might want to locate and analyze unusually effective performers for a number of reasons:

To identify for promotion or use. Personnel policies often stress the discovery

*An early version of this paper was presented at the Karachi Chapter of the Pakistan Statistical Association in August 1975. Many colleagues have lent me ideas and stimulated my own. Gus Haggstrom and Frederick Mosteller, in particular, took extraordinary pains with a preliminary draft, leading to major improvements but perhaps not as many as they hoped. Henry Acland, William Fairley, David Hoaglin, Robert Hogg, Christopher Jencks, William Kruskal, Austin Swanson, and Henri Theil also made many helpful suggestions. I must report that some of their objections remain unsatisfied, and the usual caveat protecting these courteous people from further responsibility is, of course, in order.

of particularly capable executives for rapid promotion. Scientists who breed wheat or test serums frequently assess an enormous variety of possibilities, ignoring the average effect of them all and searching for the rare outlier that works. One may infer from the work of Ithiel de Sola Pool that the training and funding of unusually effective scientists may be a key to scientific productivity (3). In education, National Merit Scholarships and teacher-of-the-year awards are examples of this reason for finding the best.

To use the unusual as a guide to the usual. Freud studied neurotics and psychotics partly because they provided extreme manifestations of psychic processes common to all. That which is difficult to study in the average case may be easier to analyze in the extreme. "From this point of view," writes Claude Levi-Strauss, "the key myth is interesting not because it is typical, but rather because of its irregular position within the group. It so happens that this particular myth raises problems of interpretation that are especially likely to stimulate reflection" (4). The unusually successful (or unsuccessful) school may provide a clearer picture of processes operating to a lesser extent elsewhere.

To avoid the oversimplification arising from the analysis of averages. Acting as if the mean (or another measure of the central tendency) provides the whole picture is not a malady confined to education (5). In investment policy, for example, one must go by the average rate of return and consider the prospect of unusually large gains or losses (6). In a critique of research on race, Ginsburg and Laughlin state, "A measure of central tendency with respect to a behavioral attribute of a genetically variable group provides very little useful information" (7). On a related subject, Jerry Hirsch (8) goes even further:

> I know of nothing that has contributed more to impose the typological way of thought on, and perpetuates it in, present-day psychology than the feedback from these methods for describing observations in terms of group averages.

Differences between groups that are small on average may be large above or below certain levels of exceptional performance (9); or, as in the case of sexual differences, there may be identical means but differences in the tails (10). Looking at the extremes in education may help us to avoid simplistic conclusions and advance our analysis of what is happening.

To imitate. Diogenes looked for an honest man to emulate; educators have looked for unusually effective pedagogical programs to imitate. That neither quest was successful (11) is an important comment about the state of the world, but also perhaps about the techniques used in the search. What statistical tools might be utilized to find the best? (12)

Simple Ways to Find the Best

The appropriate definition of "exeptional" will depend, in part, on the purpose of the evaluation. No one definition and no one statistical test will be appropriate for all purposes; and for some purposes, we may wish to use a variety of definitions and tests.

Suppose you are the evaluation officer of a school district. As part of your job, you wish to look for unusually effective schools. Let us assume that you have a performance measure X on each of the N high schools in your district. (Later we will consider the appropriateness and construction of such a measure.)

If you simply wanted to identify the K best schools, you would rank the schools by X and count down K from the top.

If you wanted to identify the schools that had scores greater than one sample standard deviation s above the average score $\bar{X}$, you would compute $\bar{X}$ and s and set off those schools with scores greater than $\bar{X} + s$.

Such tasks are not taxing. But other ways of defining your evaluative problem can cause trouble. Two deserve emphasis.

First, there is the problem of random variation. In any group, random variation assures that there will always be some set of K best schools. Similarly, some schools will have scores greater than one standard deviation above the mean. How can you tell if a school's score is really different or simply a random fluctuation?

Second, there is the problem of isolating the effectiveness of schooling. Almost any educational performance measure will be affected by differences in students' socioeconomic and genetic backgrounds, over which school policies have no control. To evaluate the effect of educational policies themselves, you must statistically or experimentally control for those nonschool factors that influence performance. In practice, this often involves the use of multiple-regression analysis, the analysis of variance, and similar techniques. How can exceptional performance be discovered when both random variation and extraneous variables affect performance?

RANDOM VARIATION: THE CASE OF A SINGLE SAMPLE

Statisticians have long tried to distinguish between random events and outliers. It is true that simply by looking at a batch of numbers one cannot tell why a particular observation is large or small compared with the rest. But one can say, imprecisely but helpfully, that "an outlying observation is one that does not fit in with the pattern of the remaining observations" (13). Then the task is how to measure "fitting in with the pattern."

In much statistical work, the search for outliers has the motive of removing from a sample those wild observations that would distort sample statistics. It has been estimated that in engineering applications, for example, about 10 percent of measured observations are outliers--sometimes caused by measurement error, sometimes by a breakdown in experimental conditions, sometimes by a "different" experimental subject (for example, a cartridge that was overloaded with gunpowder),

and occasionally by random fluctuation. Especially when the sample size is small, outliers can severely hamper the efficiency of estimators of central tendency (14) or spread. So statisticians have tried to weed them out (or cut them down to size) so that the usual summary statistics would retain their validity and power.

It may be helpful to point out three generations in the analysis of outliers, as Tukey has done for estimators of location (15). Although all three generations have their usefulness, third-generation methods are perhaps the most generally applicable but the least generally known. In order to appreciate what is new and important about these methods, it is worthwhile to examine their forerunners.

First-Generation Methods

The first generation of analysis set up the problem this way: Assume that school scores can be modeled as a sample from a normally distributed population, except perhaps for the K highest scoring schools. Can we say with some specified degree of confidence that those K schools are not drawn from the same normal distribution?

For K=1, you are examining the best of the N schools, and various statistical tests have been provided. Simulation experiments have shown that the following simple test performs as well as or better than the others (16). Put the largest score X_M in studentized form--that is, the number of sample standard deviations it falls above the sample mean (17). ($T_M = \frac{X_M - \bar{X}}{s}$.) Then compare this score with the critical value $T_{\alpha, N}$, where there is an α-percent likelihood that the largest studentized value of a sample of size N from a normal population would fall above $T_{\alpha, N}$. If $T_M > T_{\alpha, N}$, then you reject at the α-significance level the hypothesis that X_M is an item of the random sample from a normal distribution and accept the alternative hypothesis that X_M is an outlier.

Even in this relatively simple case, there are statistical problems. Only one is noted here, because it carries a more general lesson. If there are in reality two outliers and not just one, both $\bar{X}$ and s may be increased to such an extent that the test does not identify either as an outlier. With regard to the detection of outliers, this has been called the "masking effect" (18). But the more general problem might well be called a sort of "Catch 22."

The catch is that what is unusual can only be defined in terms of what is usual. But if the usual has to be gauged from sample statistics, then it, in turn, is a function of the values of unusual observations. The sample mean is greatly affected by outliers, and the sample standard deviation even more so. How can we define the usual without its being affected too much by unusual observations?

Second-Generation Methods

The unusual affects our definition of the usual only if we let it. Second-generation data analysts defined the usual in such a way that outliers would have very little effect--for example, they used the median instead of the mean. They spurned

the assumption of normality. In fact, they assumed so little about underlying populations that their tests became known as "distribution free."

One distribution-free test for outliers poses the problem this way: Suppose students of the N schools can be thought of as samples from each of n continuous populations. The null hypothesis is that all the populations are identical except perhaps for one outlier that has slipped to the right. To test this hypothesis, one needs a performance measure for each student, not just for each school. Then one takes all the students' scores for the district and ranks them. The rank sums R_i for each school are computed. If R_M, the rank sum of the best school, exceeds a certain (α, N) critical value, then it is said that this school is an outlier (19).

This distribution-free test (and most others) assumes that all the schools have identical populations except possibly the school with the biggest rank sum. This assumption is not a useful one for many practical evaluations. And the results of the test are still sensitive to the existence of multiple outliers.

Third-Generation Methods

Instead of blithely assuming normality as the first generation--or assuming almost nothing as the second--the third generation tries to devise definitions for the usual and the unusual that are robust in the presence of nonnormal distributions but efficient even when data are normally distributed.

One principal characteristic of third-generation analysts is their disdain for traditional optimality properties and exact tests (20). Tukey likens the data analyst to a detective rather than a judge: The job is to look for clues for further examination, not to pass final judgment or to derive exact confidence intervals. The third-generation philosophy is evident in the multiplicity of techniques available, the advice to be flexible and adaptive, depending on particular situations and a pragmatic definition of success: "If it works well most of the time, use it."

When considering a sample of observations, the first order of business is to transform the data so as to eliminate unnecessarily straggling tails and to induce symmetry. Logarithms and square roots are commonly used. The choice of transformation may or may not be guided by theory; the choice is usually one that the data analyst believes from experience (and from the data at hand) will lead to a fruitful exploration.

The next step is to define what is unusual. One third-generation method proceeds as follows: Roughly speaking, compute the inter-quartile range $R=Q_3-Q_1$ and consider the interval from Q_1-R to Q_3+R. Call all values beyond these limits "outside." Create new outer limits $Q_1-3/2R$ and $Q_3+3/2R$, and call all values beyond these limits "detached." In well-behaved data, about one-twentieth of the observations will be outside and one-hundreth detached.

This procedure identifies two sorts of outliers, not just one. The effort may go still further, with the use of two kinds of "skipping procedures" that Tukey extols as "more nearly the full flower of third-generation techniques than their unskipped relatives" (21). In these iterative procedures, either outside ("s-skipping") or detached ("t-skipping") values are set aside, and the hinges and midspreads are recomputed. New outside and detached values are then defined (if any now exist) and are set aside. This process of skipping is continued until an iteration discovers no new outside or detached values (22).

Such meth ..s have several advantages. If more than one outlier exists, they are more readily spotted. Exceptional performers can be found without assuming a particular underlying distribution.

The primary disadvantage is unfamiliarity. Outliers are defined by the test rather than by a more intuitive (or habitual) appeal to underlying normality or to the assumptions of the nonparametric slippage test. The method does not lead to a confidence interval or an exact test. The third-generation analyst may counter that the appeal of the usual assumptions and exact tests is more than outweighted by the fact that real data sets do not behave as assumed. For some purposes, methods based on the assumptions of normality are fairly robust; for others, such as the estimation of the center of a symmetric distribution or the discovery of outliers, they are not.

CONTROL VARIABLES: THE CASE OF REGRESSION ANALYSIS

When many variables affect an outcome, it is difficult to isolate the effect of just one. It can be even more difficult to find an outlier. The problem of defining what is usual and unusual in a multivariate situation is sure to tax statisticians for many years to come. In this section, we shall briefly consider some methods for the discovery of unusual performers in a multiple-regression context.

Imagine you are the evaluation officer in a school district whose measure of school performance is average cognitive achievement (23). You realize that achievement scores are greatly affected by differences among schools in students' socioeconomic and genetic endowments. You wish to evaluate each school on the average achievement score, given its students' nonschool backgrounds. Following a common procedure in educational evaluation, you may decide to control for nonschool factors using regression analysis (24). The difference between a school's actual score and the score predicted for it by the regression equation might then be used as a measure of the school's average achievement, given its students' nonschool backgrounds (25).

Suppose for the moment that you have a perfectly specified model, and all the usual assumptions of ordinary least squares (OLS) are fulfilled. (In other words, adopt the usual first-generation assumptions.) You then may think that the residuals from your regression equation can be used as the single sample of Section 2: Only school effects and random variation are present. Can you go ahead and apply the methods of Section 2 in order to find outliers?

Residuals with Nonconstant Variance

Even under these most favorable of assumptions, there are problems. The fact is, that even when the error terms are homoscedastic and uncorrelated, the residuals are not (26). For different schools, the residual measure of school performance will have different mixtures of school effects and random error. This fact means that test statistics which are based on t distributions (27) are no longer valid, since these statistics do not, in general, have the same distributions when there are correlated random variables with differing variances.

What is to be done? One idea is to transform the residuals into BLUS residuals--a best linear unbiased set of residuals with the additional desirable quality of a scalar covariance matrix. Unfortunately, if n is the number of observations and K the number of coefficients adjusted, only n-K BLUS residuals can be computed (28). The problem is to choose with K observations should not have BLUS residuals computed, bearing in mind that one wants to be sure the potential outliers are included. As in BLUS tests for serial correlation and heteroscedasticity, one makes the choice of the K observations to be dropped dependent on the values of the original residuals, which, in general, makes the covariance matrix of the BLUS residuals different from $\sigma^2 I$. Henri Theil suggests that this effect may be kept to a minimum by ranking the observations according to algebraically increasing values of the original residuals and dropping "appropriately spaced" observations. For example, with n=14 and K=4, one may drop the third, sixth, ninth, and twelfth observations (29). Then T_M may be applied to the BLUS residuals.

Another idea is to divide the residual by its standard error in the formula for T_M, not by the standard error of the regression equation (30). Intuitively, this would standardize the residuals in such a way that they would be comparable, so that one could treat the residuals as a single sample of measures of school effectiveness, given nonschool factors.

But then how are the critical values for the new test statistic $T_M^* = \frac{X_M - \bar{X}}{s_M}$ to be calculated? There is no exact test statistic that is applicable to all possible configurations of the X matrix. In other statistical tests on residuals, the particular configurations of the x-values can make an important difference (31). However, according to one simulation study, "...when testing for a single outlier in a simple linear regression, the effect of the arrangement of the x's is negligible, and the critical values may be obtained from Grubbs' [table]" (32). As P. Prescott (33) puts it:

> These results suggest that quite close approximations to these critical values could be obtained by assuming that the variances of the residuals are reasonably constant and using the average value of these variances in the development of the percentage points of the test statistic.

Catch 22 Revisited

Third-generation analysts might well criticize all of the above as right answers to the wrong question. Outliers can severely affect the OLS regression line, since by minimizing the sum of squared errors, the line tilts and shifts to try to make the outliers disappear. Since the usual is greatly affected by the unusual, the catch discussed earlier implies that the identification of the unusual is itself affected. In the case of OLS, there will be fewer extreme outliers than there should be. As William Fairley puts it, "OLS east outliers."

Especially in a multiple-regression context, outlying observations can affect OLS coefficients to such an extent that such points are not plainly visible as outliers, or even as the largest residuals (34).

Third-generation data analysts therefore propose various robust fitting techniques that will not let outliers affect the coefficients too much. As a result, residuals from the robust line will show up outliers more readily. But, as in the case of third-generation techniques applied to a single sample, there will not be any neat test statistic or upper bound to say with 100-α percent confidence that an outlier has shown up.

A great deal of recent work examines different methods of robust regression. Two proposed methods may be mentioned briefly here (35).

First, one may fit the line by minimizing $\sum (Y_i - \alpha_1 - \alpha_2 x_i)^P$ for other values besides p=2, the OLS method. Some Monte Carlo experiments with normal distributions that have been contaminated with varying small percentages of outliers suggest that p=1.5 is a good choice, although even lower values may be better if the contamination is serious (36).

A second method is adaptive: Make a robust fit and take the next step, depending on the residuals. One such technique of many proposed (37) is to fit the equation with p=1, to trim off a certain number of points with large residuals (38), and go back and use OLS to fit the original data without the trimmed observations. Reintroduce the trimmed values into the OLS equation and compute their new residuals.

After either fitting method is applied, third-generation techniques for outliers in single samples can be applied to the residuals.

APPLICATIONS TO EDUCATIONAL EVALUATION

How can the statistical techniques reviewed above be related to the real-life problems of educational evaluation? Several characteristics of most educational data sets, and of the purposes of educational evaluation, should be kept in mind.

1. The Need for Control Variables. Apart from direct experiments, most large-scale educational evaluations must rely on statistical controls for the many noneducational factors that affect student performance. This means that, in most cases, the search for outliers will take place in the context of regression analysis. (39).

2. The Lack of a Model for School Effects. Unfortunately, there is not now (and perhaps there may never be) a convincing model of what measurable school inputs should be combined in what way to gauge schools' effects on student performance. In my opinion, there is no believable production function to use for the indirect statistical estimation of the effectiveness of policy variables. As a result, estimates of overall school (or treatment) effects must usually be based on the residuals left over after nonschool factors are held constant. This situation also holds in many other areas of public policy: for the evaluation of fire departments and child-care programs, of collective farms and army units.

3. The Lack of a Model for Nonschool Effects. Measures of nonschool factors such as students' socioeconomic backgrounds and innate endowments are really incomplete and inexact proxies for the variables one would wish to hold constant. There will be many effects on residual variation besides school effects and random disturbance. To mention a few: misspecification of nonschool variables, omitted variables, errors in the variables, multicollinearity (40). And this fact, in turn, implies that locating statistical outliers may not locate unusually effective schools. One may simply identify schools on which these other sources of variability have their most pronounced effect.

4. The Existence of Multiple Performance Measures. No single metric can be used to gauge a school's effectiveness completely. Even for a given objective such as increasing students' cognitive achievement, one may care about many statistics of a school's scores besides the mean (41). And an outlying observation along one dimension of performance may be merely ordinary along another.

In such cases, one can pursue several courses. First, one may simply search for unusual performers along each dimension of performance considered separately. Some schools may turn out to be outliers along mean reading scores while other schools are exceptional along some measure of affective growth.

An alternative course is to see if some schools are outliers over all measures. One might calculate the number of measures over which each school was an outlier, or the number of times each school was above some threshhold of outstanding performance (say, one standard deviation above the mean). These numbers could simply be ordered--thus, the best school would be the one with the largest number of "times above"--or they could be compared with the numbers expected if all measures were independent (42). (See point 6 below.)

5. The Existence of Different Objective Functions and Production Functions. In a decentralized educational system, schools try to attain different goals or they weight the same goals differently. One would need a specification of each school's objective function--and production function--to judge how effectively a school was pursuing its chosen objectives, but such specification

is, practically speaking, unobtainable. This means that any one method of evaluating an unusually effective school would not define effectiveness as some (or many) schools would.

6. The Existence of Multiple Observations. One frequently is able to examine the performance of the same schools over a number of years and at several grades within a given year. This fact enables the analyst to distinguish random fluctuations from different school effects with more confidence. A number of methods can be used to gauge whether, over all years or all grades, a school is unusually effective compared with the rest (43).

Suppose there are n observations on each school (say, over n years). For each year i, each school will receive a residual score x_i. Residual scores may then be standardized by dividing them by the standard error of the relevant regression equations so that for school k the standardized residual vector $x_k = (x_{k1}/s_1 \ldots x_{kn}/s_n)$ is formed.

One method for deciding whether school k is unusually effective is to compare the length of x_k against the corresponding lengths for other schools (44). This technique, which can only be used for schools with all positive residual scores, has the disadvantage of being sensitive to a single large score (45).

A second method is to use the Mahalanobis distance (46). But again a single large score can give a school a high overall score.

A third method averages the n scores in x_k. Weighted averages might also be considered. Averages, however, are still sensitive to large values for a single score.

A fourth method would count the number of times that a school had individual scores over some threshhold (in the Klitgaard-Hall study, over one standard deviation above the mean). One variation of the fourth method would count how many times out of n chances a given school had an "outlying" score. Schools fallin below the threshhold are not penalized severely, even if they have large negative scores (47).

A fifth method is based on ranks. For example, for each year, one may rank the school's residual score among the rest of the schools and then calculate the mean rank over n years. This technique is relatively insensitive to a single large score, yet to some extent it does penalize schools with very low scores.

Whichever method is chosen, tests can be devised that compare a school's score to the one expected if the individual scores were independent. Under the fourth method, for example, the actual and expected numbers of times above the threshhold can be computed and subjected to a chi-square test (48). Under some scoring methods, one can treat the resulting distribution of scores as a single sample and use the techniques described above to look for the best.

CONCLUDING REMARKS

In evaluating educational programs, as in many other areas of public policy, one confronts a number of statistical problems. There is no simple measure of effectiveness. Neither does one have a believable, universal production function to use for a sophisticated estimation of school and policy effects. Therefore, relative effectiveness -- performance compared to other schools with similar students--can usually only be measured by the size of a school's residual after controlling for nonschool factors. But the relevant nonschool factors are usually not well specified or accurately measured. Consequently, there is no guarantee that residual measures of effectiveness will comprise only the effects of different schools plus random variation; and, consequently, one cannot be sure that an outlier discovered with any of the methods discussed above is truly an unusually effective performer.

This is chastening news, but it need not paralyze the data analyst. Looking for the best may be a tentative and uncertain business, but it is also a useful one. Exceptional performers may embody techniques that are copiable elsewhere; they may offer clues to the understanding of little-fathomed processes operating throughout the system; and they help to overcome simplistic generalizations based on group averages. Looking for exceptions should be a part of all statistical evaluations of public policies. Although a statistical search for unusual performers can only be a prelude to detailed case studies and not a substitute for them, it helps the scholar and the policy maker to know where to focus that attention.

It has been emphasized in this paper that any statistical definition of the unusual depends on what one defines as usual, and vice versa. Defining such terms is not trivial (49). It is often assumed that the usual is not much affected by the unusual--for example, by positing a random sample from a normal distribution. But with real data sets, such rarefied first-generation assumptions are often unhelpful. More robust, flexible, and inexact techniques are often advisable, both in defining what is commonplace and in looking for the best.

NOTES

1. For compilations of recent work, see Robert V. Hogg, "Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory," Journal of the American Statistical Association, Vol. 69, December 1974; and Peter J. Huber, "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," Annals of Statistics, Vol. 1, September 1973.

2. Or it may be an unusually ineffective school, which may also carry important policy consequences.

3. For example, fewer than 2 percent of scientists contribute over 25 percent of the published papers in physics and chemistry, and there is evidence that the quality of their papers is excellent. Big Science, Little Science, New York, 1960.

4. C. Levi-Strauss, The Raw and the Cooked, New York: Harper & Row, 1969, p. 2.

5. Cf. Robert Klitgaard, Achievement Scores and Educational Objectives, Rand Corporation Report R-1432-NIE, January 1974 and "Going Beyond the Mean in Educational Evaluation," Public Policy, Vol. 23, Winter 1975.

6. Cf. S. C. Tsiang, "The Rationale of the Mean-Standard Deviation Analysis, Skewness Preference, and the Demand for Money," American Economic Review, Vol. 62, June 1972; and William Fairley and Henry D. Jacoby, "Investment Analysis Using the Probability Distribution of the Internal Rate of Return," Management Science, Vol. 21, August 1975.

7. Benson E. Ginsburg and William S. Laughlin, "The Distribution of Genetic Differences in Behavioral Potential in the Human Species." In Margaret Mead et al. (Ed.), Science and the Concept of Race, New York: Columbia Univer. Press, 1969, p. 29. Roland B. Dixon concurs in a remark relevant to the evaluation of schools as well as skulls: "All such contrasts [in skull forms] are blurred or concealed when the measurements are averaged, and so the series of crania may in reality be in no sense uniform, but made up of several clear-cut and radically different groups, each marked by its own specific combination of characters." Cited in Louis L. Snyder, The Idea of Racialism, Princeton, N.J.: Van Nos Reinhold Co., 1962, p. 15.

8. "Behavior-Genetic Analysis and Its Biosocial Consequences." In Robert Cancro (Ed.), Intelligence: Genetic and Environmental Influences, New York: Grune and Stratton, 1971, p. 95. Hirsch suggests paying attention to the tails of distributions.

9. "For example, schools with special curricula for the academically gifted typically find six to seven times as many white as Negro children who meet the usual criteria for admission to these programs, assuming equal numbers in the populations..."; a predictable statistical outcome with two normally distributed populations differing only about 15 IQ points on the average. Arthur R. Jensen, Educability and Group Differences, New York: Harper & Row, 1973, p. 35.

10. This generalization is especially well documented for mathematical and spatial abilities. See, for example, Eleanor E. Maccoby and Carol N. Jacklin, The Psychology of Sex Differences, Stanford, Calif., 1974, pp. 118ff. Corinne Hutt believes that males have more extreme scores along almost every trait. (See Males and Females, Middlesex, England: Penguin Books, 1972, Chapter 1.)

11. In 1972 I undertook a review of recent anecdotal literature on exceptional schools and educational techniques. The volume of such literature was enormous, and the cries of Eureka widespread; but objective evidence of success was scanty indeed. In a series of comprehensive studies of exceptional programs, Michael Wargo and his colleagues found that (1) very few programs had effects that were measurably different at the 0.05 level from the usual treatment, and (2) none of those few "successes," when examined at a later date, continued to show a significantly different effect. (See David Hawkridge et al., A Study of Exemplary Programs for the Education of the Disadvantaged, Palo Alto, Calif., 1968; and Wargo et al., Further Examination of Exemplary Programs for Educating Disadvantaged Children, Palo Alto, Calif.: American Institute for Research, 1971.)

12. This article does not pretend to be a proper review of all statistical methods used to find exceptional performers. For example, useful techniques based on scanning with dummy variables for individual schools (e.g., Potluri Rao and Roger Miller, Applied Econometrics, Belmont, Calif.: Wadsworth Publishing Co., 1971, pp. 96-77), types of normal probability plotting of residuals (e.g., John W. Tukey, "The Future of Data Analysis," Technometrics, Vol. 4, 1962, pp. 21ff; and Cuthbert Daniel and Fred S. Wood, Fitting Equations to Data, New York, 1971, Chapter 3 and passim), percentile regression lines (e.g., Hogg, "Estimates of Percentile Regression Lines Using Salary Data," Journal of the American Statistical Association, Vol. 70, March 1975), and analyses based on two-way tables (e.g., Tukey, Exploratory Data Analysis, Limited Preliminary Edition, Reading, Mass., 1970, esp. Vol. II) will not be discussed. Nor will the paper consider analogous problems of locating accident-prone drivers (e.g., Joseph Ferreira Jr., Quantitative Models for Automobile Accidents and Insurance, Washington, D.C., 197[illegible] pp. 99-105) or outstanding common stocks.

13. Wilhelmine Stefansky, "Rejecting Outliers in Factorial Designs," Technometrics, Vol. 14, 1972, p. 469.

14. An exciting demonstration of this fact, accessible and interesting to statisticians and amateurs alike, is in D. F. Andrews et al., Robust Estimates of Location, Princeton, N.J.: Princeton Univer. Press, 1972.

15. "First-generation methods acted as if all values were well-behaved. This led to using the mean for a central value--with good effects when everything was indeed well-behaved. Rather too often, however, usually where violently straying values occurred, the harvest was confusion and error.

"In over-reaction to this, second-generation methods assumed every observation to be ill-behaved, and sought as much protection from ill-behavior as possible. One result was using the median for a central value. This gave extreme protection against confusion and error, but cost somewhat more than necessary when all values happened to be well-behaved.

"The return swing of the pendulum was shorter. Third-generation methods anticipate a mixture of well-behaved and ill-behaved values. Experience teaches us it is realistic to do just this." (See Tukey, Exploratory Data Analysis, Vol. I, pp. 6-31 to 6-32.)

16. H. A. David, Order Statistics, New York, 1970, pp. 184-191.

17. Frank E. Grubbs, "Procedures for Detecting Outlying Observations in Samples," Technometrics, Vol. 11, 1969. He defines s as $\left[\frac{\Sigma(X_i - \bar{X})^2}{n-1}\right]^{\frac{1}{2}}$. Grubbs also provides tables for T , N.

18. For $K > 1$, various tests have been proposed and are reviewed in David (see note 16) who suggests the following ad hoc procedure under the assumption of normality: "Apply a certain test statistic to the sample of n. If significance is obtained, eliminate the most extreme observation and apply the same test statistic to the reduced sample of $n - 1$, adjusting the significance point to the new sample size. If significance holds again, repeat the procedure until the test statistic ceases to be significantly large" (p. 191). However, the masking problem still can occur.

19. David, op.cit., p. 178. Tables are given in R. E. Odeh, "The Distribution of the Macimum Sum of Ranks," Technometrics, Vol. 9, 1967.

20. Tukey again is quotable: "The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this: 'Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.'" ("The Future of Data Analysis," op.cit., p. 14.)

"Effective data analysis requires us to consider vague concepts, concepts that can be made definite in many ways." (Frederick Mosteller and John W. Tukey, "Data Analysis, Including Statistics," in G. Lindzey and E. Aronson (Eds.), Revised Handbook of Social Psychology, Reading, Mass., 1968, p. 95.)

"We shall not seek 'the best way' of taking this data apart. (There may well be no 'best way,' in any realistic sense.) To set aside some values but not others requires a dividing line... We saw... that the dividing lines of exploratory data analysis need not be sharp--that, in fact, it was better that they were somewhat hazy." (Tukey, Exploratory Data Analysis, op.cit., pp. 4-2, 6-2.)

21. Tukey, Exploratory Data Analysis, op.cit., Vol. I, pp. 6-32.

22. The techniques were primarily invented to give accurate estimates of location for all sorts of distributions. For example, the "s-skipped trimean" is computed when the s-skipping process stops and is roughly equal to $Q_1 + 2(\text{median}) + Q_3$. See andrews et al., op.cit,. and Tukey, loc.cit.

23. Here, as in the rest of the paper, the point does not depend on the use of cognitive achievement, or any particular performance metric, as the example.

24. For example, James S. Coleman et al., Equality of Educational Opportunity, Washington, D.C.: U.S. Dept. of Health Education & Welfare, 1966; Christopher Jencks et al., Inequality, New York, 1972; Marshall S. Smith, "Equality of Educational Opportunity: The Basic Findings Reconsidered," in Frederick Mosteller and Daniel P. Moynihan, On Equality of Education Opportunity, New York: Random House, Inc., 1972.

25. See, for example, Stephen M. Barro, "An Approach to Developing Accountability Measures for the Public Schools," Phi Delta Kappan, Vol. 52, 1970; and Henry S. Dyer, "The Measurement of Educational Opportunity," in Mosteller and Moynihan, On Equality of Education Opportunity, New York: Random House, 1972.

26. This is true except in the trivial case where the number of schools is equal to the number of regressor variables, in which case the residuals are all zeroes. A proof is given in Henri Theil, Principles of Econometrics, New York, 1971, p. 196. See also Albert J. Kinderman, "On the Distribution of the Deviations from the Mean," Sankhya, Series B, Vol. 36, 1974.

27. For example, Grubbs' statistic cited on page 7.

28. See, for example, Theil, op.cit., Chapter 5; and Theil, "The Analysis of Disturbances in Regression Analysis," Journal of the American Statistical Association, Vol. 60, 1965, pp. 1067-1079.

29. Personal communication, October 1975. In the example given, the underlined observations would not have BLUS residuals computed:

1 2 3 4 5 6 7 8 9 10 11 12 13 14

30. In simple linear regression, the standard error s_k of the kth residual is the standard error of estimate $\hat{\sigma}$ times $\left[\frac{n-1}{n} - \frac{(X_k - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]^{\frac{1}{2}}$.

31. See, for example, J. Johnston, Econometric Methods, Second Ed., London, 1972, pp. 250-258.

32. G. L. Tietjen, R. H. More, and R. J. Beckman, "Testing for a Single Outlier in Simple Linear Regression," Technometrics, Vol. 15, 1973, p. 720.

33. "An Approximate Test for Outliers in Linear Modles," Technometrics, Vol. 17, 1975, p. 130.

34. "Just a single grossly outlying observation may spoil the least squares estimate, and, moreover, outliers are much harder to spot in the regression than in the simple location case." (See Huber, op.cit., p. 915.) David Hoaglin (personal communication, October 1975) suggests that, in the design matrix $Y=X\underline{\beta} + \underline{e}$, one may look for points that are sensitive regardless of the observed value of Y_i. For example, one may examine the diagonal elements of $X(X^TX)^{-1}X^T$, which tell how close the fitted value $\hat{Y}_i$ will come to the data value Y_i. These diagonal elements must sum to the rank of X and also lie between 0 and 1. Any large elements, relative to the general run of all the diagonal elements, are warnings of possible trouble--for example, the existence of wild shots or outliers.

35. It is worth noting that most Monte Carlo work on robust techniques assumes symmetric distributions. Many real-life distributions are positively skewed. The data analyst should attempt to transform the data to symmetry by the use of logs and roots in order to increase the robustness of his regression methods. (See Tukey, Exploratory Data Analysis, op.cit., and G.E.P. Box and D. R. Cox, "An Analysis of Transformations," Journal of the Royal Statistical Society, Series B, Vol. 26, 1964.)

36. See Alan B. Forsythe, "Robust Estimation of Straight Line Regression Coefficients by Minimizing pth Power Deviations," Technometrics, Vol. 14, 1972.

37. See Hogg, op.cit., pp. 915-917. The particular method given here is not explicitly mentioned in Hogg's review. See also the interesting technique of "iteratively reweighted least squares," in which the weights used in a given iteration depend on the residuals from the previous iteration. (D. F. Andrews, "A Robust Method for Multiple Linear Regression," Technometrics, Vol. 16, November 1974; and Albert E. Beaton and John W. Tukey, "The Fitting of Power Series, Meaning Polynomals, Illustrated on Band-Spectroscope Data," Technometrics, Vol. 16, May 1974.

38. More points would be trimmed for long-tailed distributions of residuals than for short-tailed distributions. No one scheme has been accepted for deciding how many points to trim in what circumstances.

39. Uncontrolled scores are also of interest, however. Analysts would be well advised to search for outliers with both uncontrolled and residual measures of achievement. (See Klitgaard, R. "Going Beyond the Mean in Educational Evaluation," op.cit., pp. 62-64.)

40. Schools with different numbers of students will also have different standard errors of estimation. Smaller schools will have larger variations in sample means of both dependent and independent variables. Such heteroscedasticity can be treated by using fitting techniques that weight schools with larger variations less, such as weighted least squares.

41. For example, one may be interested in the spread, the skewness, and the proportion above certain thresholds of achievement. (See Klitgaard R. Achievement Scores and Educational Objectives, op.cit.)

42. For a description of correlation analysis of multiple objectives, see my "Improving Educational Evaluation in a Political Setting," Rand Corporation Paper P-5184, Santa Monica, Calif., 1974. Medical literature on the idea of a normal range may be relevant here. For example, what is a normal range for blood pressure, given age, sex, weight, race, and so forth? Simply looking at ranges along conventional marginal directions would not provide a satisfactory answer.

43. For one possibility, see Robert E. Klitgaard and George R. Hall, A Statistical Search for Unusually Effective Schools, Rand Corporation Report R-1210-CC/RC, Santa Monica, Calif., 1973, pp. 24-27, 33-38; and Klitgaard and Hall, "Are There Unusually Effective Schools?" Journal of Human Resources, winter 1975, Vol. 10. Much of the following is based on an unpublished memorandum from, and personal communication with, Gus Haggstrom.

44. where the length $||x_k|| = (\Sigma x_{ki}^2)^{\frac{1}{2}}$

45. Thus, for n= 4, a school with x_k = (3,0,0,0) would have a higher score than one with (1,1,1,1).

46. The Mahalanobis distance for X_k is equal to $X_k S^{-1} X_k'$, where S is the sample covariance matrix, $\Sigma X_k' X_k / N$.

47. Thus, schools with standardized residual vectors of (2, 2, 0.9, 0.9) and (2, 2, -3, -3) are both given scores of 2 under the Klitgaard-Hall scoring method.

48. Klitgaard and Hall, A Statistical Search for Unusually Effective Schools, op.cit. and Klitgaard and Hall, "Are There Unusually Effective Schools?", op.cit.

49. Definitions will depend on the purpose of the evaluation and the data at hand. The classifying problem faced by psychopathologists is germane. "A source of difficulty may lie in the definition of what is psychologically abnormal... Several investigators.. have stressed the inappropriateness of discussing diagnosis in the abstract, pointing out that such a diagnosis should center around the question of 'diagnosis for what?' Indeed, a diagnostic system cannot be described as 'true' or 'false'..." (See Edward Zigler and Leslie Phillips, "Psychiatric Diagnosis: A Critique," in James O. Palmer and Michael J. Goldstein, Eds., Perspectives in Psychopathology, Los Angeles, 1966, pp. 14, 15.)