DOCUMENT RESUME

ED 096 331 TM 003 924

AUTHOR	Klitgaard, Robert E.
TITLE	Going Beyond the Mean in Educational Evaluation. Paper No. 5184.
INSTITUTION	Rand Corp., Santa Monica, Calif.
REPORT NO	P-5184
PUB DATE	Mar 74
NOTE	38p.; For related documents, see TM 003 350 and 845
EDRS PRICE	MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS	*Academic Achievement; *Educational Assessment; *Evaluation Techniques; Multiple Regression Analysis; Program Effectiveness; *Statistical Analysis

ABSTRACT

Evaluations in education often throw away important information because of a penchant for averages. Multiple regression techniques are used to estimate the average effect of policies across schools, and usually school performance is represented by the average score of its students on an achievement test. The author suggests some ways of broadening educational evaluations: (1) to consider "outliers," or exceptional performers among schools; and especially, (2) to consider other statistics of a school's distribution of scores besides the mean, which have an intuitive link to ill-defined but still meaningful educational objectives like equality, mobility, success with exceptional children, and attainment of certain minimum levels of skills. Tables reflecting the author's research are included. (Author/SE)

ED 096331

U S DEPARTMENT OF HEALTH.
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

GOING BEYOND THE MEAN IN EDUCATIONAL EVALUATION

Robert E. Klitgaard

March 1974

P-5184

TM 003 924

The Rand Paper Series

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

The Rand Corporation
Santa Monica, California 90406

GOING BEYOND THE MEAN IN EDUCATIONAL EVALUATION*

Robert E. Klitgaard

The Rand Corporation, Santa Monica, California

BEST COPY AVAILABLE

Evaluations in education, as elsewhere, often throw away important information because of a penchant for averages. Multiple regression techniques often are used to estimate the average effect of policies across schools. And usually the statistic of school performance is the average score of its students, say on an achievement test. In this paper I suggest some ways of broadening educational evaluations: first, to consider "outliers," or exceptional performers among schools; and second, to consider other statistics of a school's[1] distribution of scores besides the mean, which have an intuitive link to ill-defined but still meaningful educational objectives like equality, mobility, success with exceptional children, and attainment of certain minimum levels of skills. Although I confine my remarks here to the domain of education, many apply to other policy areas as well.

Averages are pleasant to work with, being easily computable and often effective estimators of the central tendency of a distribution. In evaluations of public education, researchers have been disheartened to learn that the average effect of variations in school policies on school average scores is not consistently and importantly large, once

*This article is based on two larger studies. A Statistical Search for Unusually Effective Schools (with George R. Hall), The Rand Corporation, R-1210-CC/RC, March 1973, and Achievement Scores and Educational Objectives, The Rand Corporation, R-1217-NIE, January, 1974. I am grateful to the Carnegie Corporation, The National Institute of Education, and The Rand Corporation for support; to George Hall for inspiration; and to Franklin Berger, Theodore Donaldson, Gus Haggstrom, Richard Light, and Richard Zeckhauser for advice and assistance. The usual caveat protecting these individuals and institutions from further responsibility is, of course, in order.

[1] The unit of analysis might not be schools but districts, programs, counties, and so forth. For simplicity I shall assume in what follows that the relevant unit is the school.

the students' socioeconomic characteristics are held constant.[1] As a result of these disappointing findings, educators have lashed out, alternatively or simultaneously, at achievement test measures, at public schools in general, at insufficient levels of funding or at too much funding.

But perhaps their ire should first be directed at the evaluators' penchant for averages. Even if, on average, school policies do not seem to greatly affect measurable student performance, might there not be some schools that are *exceptions* to the insignificant regression coefficients? And even if policies do not affect the schools' *average* achievement scores, might they not affect the intraschool distribution of scores--or the scores of some subset of students--in interesting and important ways?

These questions have significant policy implications If unusually effective schools can be identified, even if they are rare there is hope that their superior performance can be replicated elsewhere in the educational system. (And if no exceptional schools exist, we may have to consider alternatives radically different from current dissemination and diffusion policies--even to consider substantial changes in educational expenditures, or overhauling the entire system.) If alternative policies turn out to affect the spread of a school's scores, or perhaps the scores of gifted or retarded children, even if such effects "wash out" when we look at average scores, they may be very important for policy.

Searching for Unusually Effective Schools

Suppose one looks at school mean scores and asks whether, after controlling for nonschool factors (like socioeconomic status, geographical variables, and so forth), some schools consistently have much higher "value-added" than others. That is, are some schools consistently above the regression line that relates the nonschool factors to achievement scores? Might they be called unusually effective schools?

To find out, I looked at four large sets of achievement data, including Michigan (1969-1971, grades 4 and 7), New York City (1967-1971, grades 2 through 6), Project Talent (1960, grades 9 and 12, national

[1] Averch *et al.* (1972); Jencks *et al.* (1972); Mosteller and Moynihan (1972).

sample), and New York State districts (1969-1971, grades 3 and 6). Since there is no accepted model of the school policy variables that should be included to capture the schools' true effect, and since previous studies have shown that most interschool variation in mean scores is explained by variation in nonschool factors, the study controlled only for nonschool factors and assumed that all residual variation represented the school effects (and random fluctuation). The study was exploratory, aimed at finding exceptional schools if they existed; therefore, there was liberal experimentation with simple and complicated controls, using different kinds of data and different kinds of fits. If unusual schools were located, one could not definitely say whether their performance was due to school policies or not; but if no consistent overachievers were found, the result would be strong indeed. In effect the study attempted to estimate an upper limit on the probable number and magnitude of exceptional schools.

The findings have been reported in detail elsewhere.[1] In summary, evidence does exist that some schools are consistently outstanding. When such schools were found, they composed between 2 and 9 percent of the sample and were from 0.4 to 0.6 interstudent standard deviations above the achievement level expected from their nonschool factors. This increase corresponds roughly to these schools moving their standards consistently from the 50th percentile to the 65th or 70th; on some tests, this is almost a full grade level better than expectation. However, no matter how simple the control variables and even assuming that all residual variation represented the effects of school policies, no school in any data set was consistently able to raise its students' scores more than about 0.8 interstudent standard deviations.

Are these increases important? The study discovered schools that were statistically "unusual," but whether they were unusually effective is a question transcending mathematics. It depends on what one counts as important. Can the increases be attributed to school policies? This question deserves further research, preferably field studies; but it was interesting to note that, for the Michigan case, the unusually effective schools had significantly better-paid and more experienced teachers, and smaller classes, compared with the average school.

[1] Klitgaard and Hall (forthcoming).

The important lesson, I think, is that for both policy and research purposes, one must not rely solely on averages over all schools. Exceptions to the rule may be more important.

Looking Beyond the School Mean

The research just described looked only at school mean scores as the measure of success. But even if achievement scores are a useful indicator of some aspects of a student's cognitive growth, is the school mean score the right statistic to use for gauging the school's success? For the remainder of this paper, I would like to consider the methodological and statistical problems of deciding which statistics to use for evaluating schools, as well as to offer the results of several investigations of the empirical behavior of some other achievement score statistics besides the mean.

A measure may be useful for assessing an individual's welfare, yet the mean of that measure over a group of individuals may be quite unsatisfactory as an evaluator of the group's welfare. To show how this applies to a more familiar case than achievement scores, consider the way one evaluates income distributions. Suppose a person's economic assets form a satisfactory measure of his welfare, either because there are no other objectives than economic ones, or because a uniform metric of willingness to pay can translate other types of objectives into an economic measure (under stringent conditions that can be considered met), or because we are concerned for the moment with his economic welfare and our appreciation of that is independent of other dimensions of welfare. Suppose income is the metric for individuals, and the desire is to evaluate the welfare of a group--say, a country. What statistics are appropriate? Most people would maintain that the national average income would not be the only statistic of interest. To be sure, per capita income is widely used to rank nations' economic development and to indicate secular trends. But no description of a nation's economic welfare would be complete without some measure of the distribution of income--its dispersion among rich and poor.

Other statistics of income dispersion might be of importance in evaluating a nation's economic situation. The relative wealth of particular groups--racial minorities, sexes, ages, and so on--would not be captured by measures of inequality for the whole society. Yet these groups might be the targets of many national economic programs, the success of which could not be gauged using the national average or some index of national income distribution.

Many assessments of economic well-being also concern themselves with poverty, usually defined using a threshold below which a citizen is called poor. Generally, the mean and the dispersion alone do not reflect this concern: The statistic of interest is the proportion of the population that falls below the poverty line, whether the line is defined absolutely or relatively. Economic policies that combat poverty would be poorly evaluated using only per capita income figures or changes in the Gini index.

Educational evaluations should be similarly informed about aspects of school success beyond the average score. School policies are also concerned about equality of outcomes, success with fast and slow learners, students from underprivileged backgrounds, mobility and educational opportunity, and certain minimum levels of attainment. Judging schools only on the basis of average scores overlooks all these objectives.

Supposing we agree to go beyond just the mean score, two questions arise: (1) Beyond the average score along what achievement <u>measures</u>? (2) Beyond to what <u>statistics</u>?

What measures? Deciding which form of achievement score to use is not easy. In educational evaluation one is not just trying to assess the well-being of a group; one also wants to evaluate the contribution of policy-related variables of the educational systems to that well-being. For system evaluation, one might prefer a value-added or <u>residual</u> measure of achievement, not the achievement scores themselves. The reason is straightforward: pupils bring different amounts of intellectual capital to their learning experiences because of differing socioeconomic, psychological, and genetic backgrounds. Schools with superior students will tend to attain superior results, but not necessarily because of superior schooling.[1]

[1] Smith, using data from the Equality of Educational Opportunity Survey, found that only between 5.85 percent and 7.46 percent of the variation among unadjusted school mean achievement scores is <u>potentially</u> due to school effects. Cited in Jencks <u>et al</u>. (1972), p. 178.

Therefore, many writers have called for the use of residual achievement scores to evaluate public education. Only by taking the students' varying nonschool background factors into account, they argue, can the differences between school scores be linked to the quality of the education provided.

Residual scores also have their opponents. There are a host of statistical problems, not least of which is choosing the appropriate control variables. At best socioeconomic measures are proxies for the background factors one wishes to hold constant across schools, and the predictive power of various controls may differ from community to community, making residual scores difficult to interpret.[1] Some argue that residual scores computed from school-level data are subject to computational unreliability.[2] Even working with individual residual scores is subject to statistical errors of many kinds.[3] If there is

[1] These problems are often recognized by advocates, but usually left unresolved; see, for example, Barro (1970), pp. 203-205 Dyer (1972), p. 526 concludes cheerfully:

> Anyone who examines closely the method I am proposing for assessing the educational opportunities provided by schools will find plenty of problems in it, some theoretical or technical and some practical. There is no space here to discuss these problems, but I am convinced that, possibly with some modification of the basic model, they can be solved.

For a less sanguine view, see Cronbach and Furby (1970).

[2] Dyer, Linn, and Patton (1969), implicitly assuming that separate regressions used to control individual scores and school scores for background factors were free from error, found that school-level residuals had undesirably low correlations with aggregated individual-level residuals for the same schools.

[3] Residual variation could arise from other causes than differences in school effectiveness: imperfection in measurement, misspecification of background factors, omitted variables, poor choice of fitting technique, incomplete data, regression toward the mean, and the combined random fluctuations involved in all the regressor variables.

multicollinearity between school variables and nonschool background factors, further uncertainty is introduced into the estimation of school effects.[1]

A non-statistical, normative problem also attends the use of residual scores. Evaluating with residual scores implies that the regression line (relating background factors to achievement) is accepted as the normative baseline from which to judge policy. To some educators, the fact that the regression line indicates differences in achievement across economic classes, geographical areas, and racial groups is part of the problem and is itself an indicator of poor performance by the educational system. Some educators have maintained that using residual scores endorses existing inequalities as the proper frame of reference for evaluation.

The choice of measures may depend on the choice of problems one wishes to analyze. To evaluate _cost-benefit_ aspects of education--to compare the educational dollar's productivity with a dollar for defense, housing, or tax refunds--one may prefer an absolute achievement measure. However, for _cost-effectiveness_ questions--to compare one school or educational practice with another--a residual measure may be better.

There may be no need to be exclusive. Both measures are useful, and both convey different kinds of information about a school's performance. The wisest strategy, then, might be to use both unadjusted achievement data and achievement residuals.

The mean is a useful summary statistic of a school's performance under certain circumstances. But using only the mean for evaluation both throws away information and makes assumptions that are probably untenable. Using the mean for evaluation implies:

(a) An increase in an achievement score of a given magnitude is valued equivalently, no matter where on the achievement scale it occurs. (A gain from 25 to 30 is just the same as a gain

[1] Given multicollinearity, the significance of each affected variable will be difficult to interpret. Also, if the amount of multicollinearity varies from regression to regression, not only will significance tests be difficult, but techniques for partitioning shared variance will give different answers. See Mayeske _et al_. (1969) and Craeger (1971).

from 65 to 70, for example.) But the assumption is false if we care particularly about the attainment of certain basic skills, or if high scores are very desirable. Where educational policy does not equally value equal-sized gains on a standardized achievement test, the mean will not accurately reflect educational objectives.

(b) All students are valued equally (since the arithmetic mean adds all students' scores in an unweighted fashion, dividing by the total number of students). But educational policy may attach greater weight to academic gains among certain students, perhaps to overcome past disadvantages or to increase the proportion in certain academic specialties. Insofar as a policy is directed at certain types of students, the mean school score will not be adequate for evaluation.

(c) Student i's score is independent of student j's (the mean merely sums scores, without adjusting individual scores depending on the scores of others). This assumption may be false for two reasons. First, one may care about the *distribution* of scores across students: the equality of outcomes, the amount of mobility, the riskiness of educational outcomes, the tails of the distribution of scores. The mean does not communicate the distribution, just its central tendency; the analogue to income distribution is obvious. Second, if education acts as a screening device or filter for later education or for the job market, scores i and j cannot be treated as if they were independent.

Specifying Objective Functions: The Theory Versus Educational Realities

Which additional statistics should be used in evaluation? This question asks for a specification of the "objective function" that schools should have for achievement scores. An objective function is the formal link between objectives and evaluative measures. The idea behind an objective function is to assign a numerical value (utility) to every (relevant) state of the world; the decision problem is to maximize that function subject to budget and operational constraints. With such a function a school or program can be evaluated merely by examining its utility score and the costs of attaining that score.

To construct an objective function for achievement scores, three questions require answers:

(a) How does one evaluate one achievement score compared with another (or one residual score compared with another)? We may tautologically define some objective function $U_A = f(A)$, where A signifies the achievement score, or some function $U_R = g(R)$, where R signifies the residual score, but what do the functions f and g actually look like?

(b) How could U_A and U_R be combined into a single, composite objective function U_T for each student?

(c) If one is evaluating schools and not students, how could the U_{Ti} be combined for each student i into a school index?

Question (a). How does one compare scores of 35, 40 and 45? We know that 35 is five points lower than 40, and 40 five points lower than 45. But the units here are derived through some standardization process used by the testers, norming scores to some population of students. There is no necessary reason why this scale should correspond to one's evaluation of those scores. Does one equally value a five-point increase whether it is from 35 to 40 or from 40 to 45 (or from 60 to 65)? To answer this question a utility function for an individual's score is required.

Theoretically, the evaluator could construct this utility function by presenting the decisionmaker with choices between lotteries on scores. For instance, is it better for a student to have a score of 50 for sure or a 50-50 lottery on scores of 40 and 75? If you were indifferent, your utility function for the student's achievement could be suspected of being convex over that region. In the well-known von Neumann-Morgenstern fashion, a set of lottery questions could ascertain the entire function of a rational decisionmaker.[1]

[1] Von Neumann and Morgenstern (1944). See also Friedman and Savage (1948); a lucid elementary exposition is found in Raiffa (1968), Ch. 4. Roche (1971) had local educational administrators make explicit their utility functions for different kinds and levels of student achievement scores.

It is difficult to predict what utility function for achievement scores would be specified. Decisionmakers might well disagree. One answer--though in my opinion unlikely--is that in fact a five-point achievement score increase would be weighted the same whether it were from 35 to 40 or 60 to 65 or anywhere else. In such a case, U_A would be some linear function of the score, as in Fig. 1a.

Another observer might consider increases in low scores more valuable than gains in scores that are already high. If questioned in detail about his preferences for a student's scores, this observer might respond with a U_A curve like the one in Fig. 1b.

If one valued achievement gains on both the low and high ends more than those in the middle--perhaps because of an emphasis on slow learners and the gifted--a cubic utility function like Fig. 1c might be the appropriate representation.

Suppose one's educational objective were predominantly to ensure that the student achieved a score above some minimum level k--perhaps some threshold of needed cognitive skills. Achievement increases beyond k are relatively unimportant. Then Fig. 1d might be the right utility function to use for evaluation.

Clearly the shape of U_A might be many things besides linear. Different policymakers might choose different functions; different programs might want to weight achievement gains differently; and utility functions might vary for different kinds of students. Similar remarks apply for U_R: _a priori_ it seems unlikely that g(R) should be linear, and no other shape recommends itself as the obvious alternative.

Question (b). Suppose we have elicited U_A and U_R. How can we combine them into some overall utility function U_T? Theoretically, to answer this question one first assesses the interdependence of the two functions. Does our evaluation of U_A for student i depend on his residual score? That is, is the choice among lotteries on achievement scores any function of the student's residual score, or vice versa? If we hold the residual score fixed at some level R_0, do our conditional (probabilistic) preferences for the unadjusted score A depend on what fixed value R_0 is chosen, and vice versa? If _not_, then the composite utility function U_T has an additive representation:[1]

[1] Raiffa (1969).

$$U_T = U_A + U_R.$$

If our preferences for achievement scores are dependent on the student's residual score or vice versa, then U_T must be estimated in a more complicated way, by asking lottery questions among many possible achievement and residual score combinations.[1]

Question (c). Suppose U_{Ti} has been constructed for each student i. How can U_{Ti} be summed to obtain a school index of success? Once again the answer depends on the interdependence of the components to be combined. If U_{Tk} (the utility for student k) is held fixed at some level $(U_{Tk})_0$, do our conditional (probabilistic) preferences for any other U_{Ti} depend on what fixed level $(UT_k)_0$ is chosen? If not, and if the question can also be answered negatively for all U_{Ti} fixed, then U_{Ti} for all students 1,..., n are mutually preferentially independent.[2] If this independence holds, then U_T (school) can be expressed as an additive value function:

$$U_T(\text{school}) = U_{T1} + U_{T2} + \ldots + U_{Tn}.$$

In other words, if mutual preferential independence exists, evaluating a school merely involves evaluating each student and summing up the utilities over all students in the school.

Unfortunately from the point of view of analytical simplicity, such independence seems not to hold across students. As soon as distributional considerations enter--when we care about equality of outcomes, for example--then our feelings about U_{Tk} _do_ depend on the levels of the other students. Furthermore, if part of the education's value is a screening or credentialing device, then each student's scores affect the utility of his comrades' scores. Therefore, mutual preferential independence does not seem to exist. As a result, U_T(school)

[1] See Raiffa (1971) for details; Raiffa (1968, Ch. 9, Sec. 3) for an outline of the complexities.

[2] Mutual preferential independence means that the decisionmaker's substitution rate between U_{Ti} and U_{Tj} does not depend on any of the values of components other than i and j. See Raiffa (1971), pp. 74-75.

can be assessed only through a very complicated series of tradeoffs, holding each U_{Ti} fixed at different levels while assessing the remaining $U_{T(n-1)}$: a theoretically possible but operationally unpalatable task.

Using the school mean score as the evaluative statistic assumes a linear utility function and mutual preferential independence, neither of which seems true.

Turning from theory to reality, two important facts about education must be reckoned with:

(1) Local school districts (and, within districts, various interested parties) are likely to have different utility functions.

(2) Practically, it will be extremely difficult to obtain an operational specification of utility functions from educational decisionmakers.

These two propositions have serious implications for educational evaluation. Both make the methodology of utility functions less than perfectly applicable.

The first point implies that the search for a national objective function that somehow combines local preferences is futile. Consensus on education objectives will not be forthcoming--and perhaps rightly so. In a decentralized educational system, local preferences possess a certain autonomy, a certain right to be different. To evaluate all schools by the same criteria, with the same utility function, would be an error.[1]

[1] Note that the current ways of using many statistical methods to evaluate schools assume common objective functions (and production functions) among schools. Insofar as schools are trying to do different things, regression coefficients relating certain inputs to a common output may be misleading; coefficients of multiple correlation may be looking at the wrong type of variability; good schools may merely be the ones that are trying to do what one is trying to measure. Even if schools share a common objective, they will probably weight it differently in their tradeoffs among their other goals.

There still may be a justification for making evaluations according to a single objective function. Suppose, for example, that the evaluator is the federal government. A decentralized educational system does not preclude the existence of national-level spillover effects from schooling. The federal government would want to affect the local production of these effects through grants-in-aid, legal constraints, taxes, and so forth, even if not through overt control; and the federal government could evalate its success at doing so with a single national-level objective function

The second point means that, in educational evaluation, the objective is not specified in advance. The problem, in my opinion, is not that objective functions are theoretically impossible to get; the constraint is instead one of feasibility. Three problems may be mentioned: cost; the ticklish task of defining decisionmakers among the many educational officials with interests and pretensions; and if there are multiple decisionmakers, combining their objectives in a meaningful way. In practice one cannot begin with tightly defined objective functions and then deduce from them the appropriate way to use achievement measures for evaluation.

From the systems analyst's point of view, education is the worst of worlds. First, there are no well-specified objectives and they probably cannot be obtained. Second, evaluations must nonetheless be made. Third, the data are mostly restricted to achievement scores. And finally, most existing large-scale evaluations and governmental data banks use only mean scores. We know something about educational objectives--not a sufficient amount to draw curves and derive combinatorial rules, but enough to know that the present reliance on the mean is inadequate.

What should be done?[1] The situation is somewhat analogous to the one faced in evaluating a nation's economic welfare. Clearly the

that gave utility to the particular spillovers in question. This would, of course, be a very limited sort of evaluation, but perhaps this is all the federal government ought to attempt in a decentralized system.

[1] Our system analyst, an ideal type who nonetheless sometimes speaks with the same voice as more reasonable people we know, might suggest the following: "Since your decisionmakers are diverse and no mathematical algorithm can be conveniently adduced for any one or all of them, why not solve your 'statistics for evaluation' problem by giving the entire distribution of scores for each school to all the decisionmakers? Let them make up their own minds what is important." Visions of policymakers trying to examine hundreds of histograms, or having to compute residual measures according to their individual perceptions of the proper control variables, may not occur to our analyst (or, if they do, they may only cause him delight). We do not want to overwhelm decisionmakers with data. A map on a 1:1 scale is of little use. Our goal is to provide a few easily computable, informative statistics that correspond (roughly) to the policymaker's likely educational objectives and that are likely to increase his knowledge of educational outcomes.

average income statistic is not enough; clearly, too, no social welfare function has been derived from which the appropriate statistics for evaluation could be deduced. But there is a notable difference. Unlike education, national economic policy *has* employed statistics that go beyond the mean: measures of income distribution, the poverty line, and others. These statistics were not deduced from an objective function, and there is no one set of them that commands universal assent as the best and most efficient. But a number of useful statistics *have* been proposed to measure certain ill-defined although meaningful goals of economic policy. Rather than staying where we are in educational evaluation, or throwing out achievement tests altogether, perhaps we would do well to follow that example.

Statistics of Spread

Equality is an increasingly voiced goal of education. In America discussions of equality have traditionally centered on equality of opportunity: that everyone have an equal *chance* to obtain a good education, but not necessarily that everyone actually use that chance. However, many recent writers, including some of a radical bent, have emphasized equality of *outcomes* as a major educational aim. They maintain that instead of evaluating some prior notion of the opportunity schools provide--or perhaps in addition to such an investigation--the equality of the actual results should be examined.

It is not clear that the more equal the educational outcomes, the better; one's utility function might not be an increasing function of the amount of equality.[1] The central point is not that equality is preferred indefinitely but that some measure of the equality of outcomes that a school provides is helpful in a well-rounded evaluation of its effectiveness.

A school's mean score alone tells nothing about its equality of outcomes (although a comparison of school means will indicate something

[1] Despite the common usage of terms like "equality" as if they were to be maximized, there is almost surely some limit in everyone's mind--although, as Kristol (1972) points out, advocates of equality and mobility are reluctant to define optimum levels.

about equality among schools). To evaluate a school's equalizing ability, one needs to go beyond its central tendency to some estimator of the spread of the school's distribution of achievement scores.

Figure 2 shows two hypothetical distributions of achievement scores corresponding to schools A and B. Other things equal, an advocate of equality of outcomes would prefer school A because of its smaller variability, even though the school mean scores are equal.

One statistic of interest, then, is the spread of a school's uncontrolled achievement scores. Other things equal, the smaller the spread, the greater the equality of cognitive achievement outcomes.[1]

Two kinds of residual scores related to the spread can also be useful. First, suppose one is interested in comparing schools' equalizing abilities. The different degrees of equality within schools may stem from differences in nonschool background factors from school to school, rather than different equalizing effects in schools. Schools having students with more similar backgrounds can expect less variation in achievement scores. One could regress some statistic of equality of outcomes (say, the standard deviation of school scores) against various background factors to compute a predicted standard deviation for each level of the background variables. A residual score--observed standard deviation minus predicted standard deviation--could then be obtained for for each school. The smaller this residual, the greater a school's equalizing ability.

A second residual spread measure might serve as a proxy for "educational mobility," another goal of schools. Americans have long cherished the belief that education can be a powerful weapon for social advancement, without students being imprisoned by their socioeconomic backgrounds. Some recent studies, using mean achievement scores, have eroded this faith. But is the mean the right statistic to measure the effects schools have on mobility?

[1] Some educators apparently believe that larger spreads indicate superior schooling: "Every experienced teacher knows that effective teaching will increase the variance of the group being taught, and usually markedly" (Guba, 1967, p. 61).

For this mobility objective, the spread of achievement residuals may be a useful indicator. (In general, the spread of the residual scores will not be the same as the spread of the raw scores.) Given schools with equal mean residuals, the one providing greater residual variation is providing greater educational mobility. Its students have more opportunity to "succeed"--and more to "fail"--compared with other schools whose students have like socioeconomic and personal characteristics. Putting it another way, the students in a school with a larger variation of residual scores are less likely to end up where their backgrounds would have predicted.

As with equality of outcomes, it is not necessarily true that the more such "opportunity" for success and failure exists, the better. One may prefer to have less chance of failure even at the loss of some opportunity for success. In 1523 on the Isla de Gallo, Pizarro drew a line with his sword in the sand and told his men on one side lay "untold hardships and starvation, treacherous reefs and storms, bitter war and even death, but there also the golden land of the Incas" and on the other "peace, but the peace of poverty." Only 13 of the hundreds joined him on the side of possible riches. Risk preferences and distributional considerations are important in deciding how much opportunity for mobility we prefer.[1] The fact that mobility may not be indefinitely preferred does not, however, mean that the spread of residual scores is a useless measure. It is merely a reminder that "mobility" is two-directional, and that more of it, in education as elsewhere, may not be unequivocally desired.

[1] Risk preferences are important because people with higher risk aversion tend to prefer narrower distributions of outcomes to wider ones, given equal expected values.

Distributional considerations may enter if the residuals display heteroscedasticity. (Heteroscedasticity refers to nonconstant variance of residuals around the regression line.) In such cases an increase in the overall variance of a school's residuals increases the opportunities for students of certain backgrounds more than others; one cannot a priori presume that every student has the same probability of being located anywhere on the school's distribution of residuals. Therefore, which students get more opportunity becomes paramount--and this brings distributional objectives into the picture.

There are, then, three possible measures of spread that would be useful in educational evaluation: the spread of the unadjusted achievement scores, indicating equality of outcomes; the difference between the actual and expected spread of achievement scores, a proxy for the equalizing ability of schools; and the spread of the residual scores of a school's students, indicating the amount of educational mobility a school provides. Which statistic should be selected to measure spread?

There are many possible statistics of dispersion and equality. One is the variance (or its positive square root, the standard deviation). However, the variance is very sensitive to extreme values; it is not a robust estimator of spread. One estimator of spread that is less vulnerable to outliers is the interquartile range (others are given in Tukey, 1970, Vol. I, Ch. 2).

Which statistic of spread to use should depend on a careful specification of the educational objective function; but, short of this, what matters is that some such statistic be available. Further research should be devoted to selecting the best statistics of spread for education, although as in income distribution, optimality properties may not be agreed upon. With any of a number of measures of dispersion, schools could be compared cross-sectionally and over time in a useful way; the value of such statistics for evaluation should not be underestimated because of some misplaced desire for cardinal precision.

How much do schools differ in the spreads of their achievement scores? Do nonschool background factors explain differences between the spreads of schools? Is there any evidence that some schools consistently provide less variability of scores than others, holding nonschool factors constant? Since spread measures of the intraschool distribution of test scores have largely been ignored in the past, little is known about the empirical characteristics of such measures.

The following are merely preliminary investigations into the behavior of some standard deviation measures based on Michigan data for fourth and seventh grades in 1969-70 and 1970-71.[1] Since the data were

[1] The data base is described in Brown (1972) and Klitgaard and Hall (1973).

already aggregated at the school level, the "mobility" statistic, which must be based on student-level regressions, could not be computed, Only the standard deviation of unadjusted scores ("equality" statistic) and the difference between the expected and the observed standard deviation ("equalizing ability" statistic) were examined, and these two only in an exploratory fashion.

How should one expect the standard deviation statistic to behave? It is the square root of the variance, and it is similarly sensitive to extreme values in the distribution. In normal samples, the sample variance is distributed as a multiple of a chi-square variate with N-1 degrees of freedom. With N small (say, less than 10), the chi-square distribution is positively skewed; but by N = 20, the distribution is close to Gaussian. The standard deviation tends to have higher variability for smaller N; schools with fewer students tested will have a higher proportion of high and especially low standard deviations, other things equal.

In the Michigan data N (the number of students tested per grade) varied considerably from school to school (see Table 2), making school standard deviations not perfectly comparable; but since the average value of N was quite large, the analysis simply used the standard deviation without worrying about transformations. Eliminating all schools with N < 5, the average school standard deviation was about 9 and the standard deviation of the standard deviations was about 1.1 (see Table 2).[1] The distributions of school standard deviations across schools were negatively skewed.[2] This fact might well be the

[1] The achievement tests are normed to have an interstudent standard deviation of 10 and mean 50.

[2] The data cover reading and mathematics scores for fourth and seventh grades in 1969-70 and 1970-71, a total of eight sets. Not every school has both fourth and seventh grades, and not every school reported data for each possible test/grade/year combination. The skewness statistics were:

R 4 69-70 = -0.48	R 4 70-71 = -0.62
M 4 69-70 = -0.47	M 4 70-71 = -0.28
R 7 69-70 = -0.68	R 7 70-71 = -0.64
M 7 69-70 = -0.94	M 7 70-71 = -0.98

R 4 69-70 stands for the reading score for fourth grades in 1969-70; the other symbols are interpreted similarly.

result of lower variances of smaller schools. It also might indicate that some schools are trying to obtain more equality of outcomes than others, or are better at doing so than other schools with similar goals.

How do these standard deviations compare with those expected, given the different background factors among the students of different schools? To find out, a series of regressions were run, fitting the school standard deviation to a number of nonschool background factors. The best set of regressions, although still only crude and exploratory, is given in Table 1. Table 2 shows the means and standard deviations of the regressor and response variables.

The proportion of variation explained by the regression results varies rather widely, from 0.11 to 0.37. No differences seem important between the reading and mathematics regressions, although the reading scores display more heteroscedasticity as indicated by the greater significance of the μ regressor. (This difference is most striking between the fourth grade reading and mathematics scores.) SESσ has the expected positive sign on all regressions. %MIN is consistently negative, indicating that greater numbers of minority students tend to go along with the lower standard deviations, even after controlling for SES and the achievement score μ. The number of students tested N has the expected positive sign, indicating that smaller schools do tend to have smaller variability.[1]

The major finding of these regressions and the others that were tried is the <u>limited ability of background factors to predict school standard deviations</u>. This result, of course, contrasts markedly with the results of regressions on school means, where most of the variation across schools is explained by socioeconomic, racial, and regional variables. (For example, the R^2 values for simple regressions on means using the same Michigan data ranged from 0.59 to 0.78 (Klitgaard and Hall, 1973, p. 46).) One might hypothesize that the low explanatory power of background factors indicates that school policies determine standard deviations. But the low R^2 values may merely be a product of

[1] The statistical properties of the standard deviation statistic would lead one to expect smaller variances for schools with small N, even if all schools had drawn their students randomly from the same population; it also may be true that smaller schools tend to have more homogeneous student bodies, even after controlling for SESσ.

BEST COPY AVAILABLE

greater random fluctuation or purely statistical problems. This question awaits detailed investigation.

The residuals from these regressions constituted the second spread measure discussed above--a statistic purporting to indicate the equalizing ability of schools given their students' backgrounds. The distributions were slightly tighter: The standard deviations (of the standard deviations) now averaged about 1.0. Skewness was reduced, although all eight distributions are still negatively skewed.[1] Outliers remained on the left tail, but a few also showed up on the right tail now.

The extreme values on the left tail looked interesting enough to pursue. Each histogram of schools' scores (say, for a particular grade, test, and year) will show the effects of random variation as well as the effect of different schools. A thick left tail does not by itself prove that these schools with low variability are anything more than random deviates. But if the same schools show up on the left tail consistently over many grades, tests, and years, one might conclude that the phenomenon is not just a statistical fluke. Do some schools consistently record low variability, even after allowing for nonschool background factors?

To find out, the following null hypothesis was formulated: All variation of the difference between actual and expected standard deviations is a result of chance and not of school effectiveness. To test this hypothesis, some sort of "cumulative distribution" is required indicating how well schools have done over many grades, tests, and years after controlling for background factors. Then it would be possible to see if that

[1] Skewness statistics were:

R 4 69-70 = -0.20	R 4 70-71 = -0.26
M 4 69-70 = -0.14	M 4 70-71 = -0.01
R 7 69-70 = -0.36	R 7 70-71 = -0.37
M 7 69-70 = -0.16	M 7 70-71 = -0.19

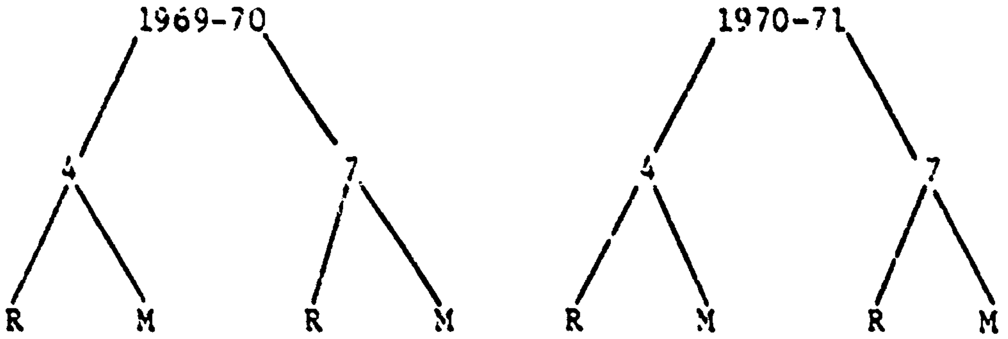
Notice especially how the mathematics scores have become less skewed.

BEST COPY AVAILABLE

distribution differed significantly from a theoretical distribution obtained by treating all the individual distributions of residuals as statistically independent.

As a proxy for this cumulative distribution, each school in a given grade, test, and year distribution was assigned a score of one if it was more than one standard error below the mean and a score of zero otherwise. Each school's totals were added up over all distributions, and a chi-square test was used to see whether some schools were consistently below one standard error more than chance would predict. The results appear in Table 3.[1]

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



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

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

The actual probabilities for each pair of tests are:

BEST COPY AVAILABLE

There is evidence in Table 3 that some schools consistently have a greater equalizing effect on their students' achievement scores than chance alone would predict. The schools that were consistently below average did tend to be quite a distance below each time. For example, the ten schools that were below one standard deviation at least five out of eight times averaged about 1.6σ below each time. Since the standard errors were about one test score point and the interstudent σ = 10, these ten schools were reducing the variability of their students' scores about 1/6 of the interstudent variation compared with the average school. On the fourth grade Iowa reading test, this would imply tightening the standard deviation of outcomes about 20-25 percent of a grade equivalent.[1]

(R_i, M_i)	P(X=0)	P(X=1)	P(X=2)
4 69-70	0.802	0.149	0.049
7 69-70	0.823	0.124	0.053
4 70-71	0.804	0.139	0.057
7 70-71	0.845	0.102	0.052

If the school reported eight scores, it had eight chances to be below one standard deviation less than the mean; the null hypothesis is computed for four pairs of tests. If a school only had six chances, then the test is computed from three pairs; if four chances, two pairs. The chances only occurred 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, I assumed a common probability distribution P(X=0) = 0.82, P(X=1) = 0.13, P(X=2) = 0.05 for all pairs and assumed it did not matter which particular pairs happened to make up a school's set of chances.

For the chi-square approximation to be accurate in contingency tables with more than one degree of freedom, cells with small expectations must be pooled. I followed a pooling rule proposed by Yarnold (1970, p. 865):

> If the number of classes k 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/k.

[1] Lindquist and Hieronymus (1964). To give another intuitive idea of what this reduction in variability means, a 1/6 reduction in the standard deviation on most IQ tests would be 2-3 points.

BEST C... ...LABLE

It must be reemphasized that these results are only explorations. They have barely touched the surface of the important questions concerning standard deviation and other spread measures in education. How do different measures of spread behave? How important is the variability involved? How does spread relate to school and background characteristics? Perhaps this beginning can whet some appetites and suggest some directions for further study.

Statistics of Distortion

In recent years especially, educational policy has laid heavy stress on special programs for disadvantaged and gifted students. Spurred by the conviction that curricula and methods designed for the average pupil do not teach slow and fast learners efficiently, reformers have created programs for special students at an unprecedented rate. Evidently, many educators base their judgments of school quality partly on the number and sophistication of programs for different kinds of students. If educational policy is significantly directed at slow or fast students, a school's average achievement scores may be a misleading measure of its success.

Take the case of uncontrolled achievement scores. Suppose very low scores are very undesirable, very high ones extremely nice, and those around the middle more or less the same. Low achievers might be harmful to society to a far greater extent than the linear weighting of their achievement scores would indicate, while high achievers might be deemed extremely valuable. In this case, the utility function might look like the cubic function in Fig. 1c. We may be willing to let those in the middle achievement range drop a little if we can thereby move both tails of the distribution of scores to the right. For example, in Fig. 3 we may prefer school A to B, and either A or B to school C, despite equal means and variances. Distribution A has more students below the mean than B, but most are in the range where it does not matter too much; meanwhile, A's lower tail is smaller and its upper tail broader.

BEST COPY AVAILABLE

One proxy for such preferences might be the skewness of the distribution, defined as

$$E\left(\frac{(X-\mu)^3}{\sigma^3}\right).$$

Positive (negative) skewness indicates that for any specified mean and variance, the mode is likely to be smaller (larger) than the mean, the left tail "unusually" short (long), and the right tail "unusually" long (short). Increasing the positive skewness of a school's distribution of scores trades off losses around the middle of the distribution for gains in scores on both tails. Other things equal, much of educational policy probably favors positive skewness.

Similar remarks apply to the skewness of the school's distribution of **residuals**. Fit individual student scores against their nonschool background variables; compute individual residuals for each student; then aggregate those residuals by school and compute the school's skewness statistic for the distribution of residuals. Suppose that we care more about underachievers and overachievers (no matter what the score their background factors would predict). If we wish to avoid large underachievers and produce large overachievers, and if we do not care much about performances relatively near to expectation, then, other things equal, the skewness of the distribution of residuals may validly order schools according to our preferences.

Because the skewness statistic is a nonlinear functional, strictly speaking there is no von Neumann-Morgenstern utility function corresponding to its maximization. However, despite this rather ungainly feature, the skewness statistic has a history of use in econometric studies to measure exactly the phenomena relevant here: high emphasis on large positive payoffs and great displeasure at large negative ones (Tintner, 1942; Hicks, 1950; Arditti, 1967; Fisher and Hall, [illegible]). Used in regressions that also control for mean and variance, the skewness statistic--or some such measure of the distortion of the [illegible] school distribution--seems an appropriate additional measure for educational evaluation.

Once again, the precise mathematical definition of the statistic of distortion to be included is not of prime importance, nor would one prefer positive skewness indefinitely. What matters is that some indicator of distortion be available as an evaluative tool.[1] Other things equal, the more positively skewed the distribution of raw scores within a school, the better a school is doing with its slow and fast learners, although at the expense of its average students. And for individual residuals, with other things equal, the more positively skewed the distribution within a school, the better a school is doing with its under- and overachievers, although at the expense of students who perform at about the level predicted by their socioeconomic backgrounds.

Statistics of Proportions above Certain Thresholds

If some minimum level of attainment is of concern, the mean school score can easily mislead. A simple and useful measure is available: the proportion of students who score above the level in question.

A number of writers imply that certain thresholds of achievement are of the utmost concern.[2] High schools are sometimes judged by the proportion of their graduates that can read at the ninth-grade level or that go on to college, to name two quite different thresholds. In performance contracting experiments, fees often depend cn the number of students performing at or above their grade levels. For such

[1] There are problems with the skewness statistic. It is extremely sensitive to outlying values--more than the variance or the mean--and a more robust estimator might be called for. Another problem concerns the fact that one's preferences for skewness cannot be separated from one's preferences for mean and variance. Even to find a function that ranks distributions in the same order as maximizing the third moment of a distribution $E(X - \mu)^3$ involves specifying the mean and variance as well. However, with some such measure one can obtain further information that generally goes beyond the mean (which weighs all gains and losses the same no matter where on the distribution they fall) and the spread (which evaluates bigger tails on either end the same). This fact implies a lack of preferential independence among the goals relating to mean, variance, and skewness of a school's distribution: How much skewness one prefers has to depend on the level of the school's mean and variance.

[2] A lower tail threshold is implicit in the writings of Kenneth Clark, for example. Similar sentiments may be discerned in the writings of John Stuart Mill:

objectives, the proportion of students above a certain score is the best indicator of success.

As with the other statistics discussed so far, the proportion above certain thresholds has useful applications with both uncontrolled and residual scores. The proportion of students above some absolute level tells us one thing about a school; the proportion achieving above some level relative to their backgrounds, quite another. Both measures usually go beyond the information provided by means, variances, and skewness.

Some crude indications of how threshold measures behave can be gathered from data from the Yardstick Project in Cleveland, Ohio. Yardstick contracts its data analysis services to some 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 varying time spans. Thus the data base is not necessarily representative nor is it useful for longitudinal analyses. However, the Yarkstick 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.

For 72 schools separate regressions were run on school mean growth (mean score for year N minus mean score for year N-1), school mean growth for students with IQs higher than 123, and school mean growth for students with IQs lower than 93. Control variables included father's occupation and mean school IQ, among others.

Background factors do not predict success with slow and fast learners nearly as well as they predict school success with average students. For the school means, a stepwise regression yielded $R^2 = 0.55$. The other fits were very poor. In the regression on school mean growth among its students with IQ > 123, only the percentage of children in the school whose

"It may be asserted without scruple, that the aim of all intellectual training for the mass of people should be to cultivate common sense; to qualify them for forming a sound practical judgment of the circumstances by which they are surrounded. Whatever, in the intellectual department, can be superadded to this, is chiefly ornamental." (The Principles of Political Economy, Book II, Chapter XII: cited in Vaizey, 1962, p. 20).

fathers were skilled workers was significant (with a negative coefficient), and the R^2 was only 0.18. On the under 93 side, no variables reached the $F \geq 4$ significance level needed to enter the regression, and when all controls were forced into the fit, the R^2 rose only to 0.13. These results suggest, but shortcomings in data did not enable me to verify, that school variables may make more difference than background factors in determining the achievement of exceptional children, either because schools concentrate their efforts there or because schooling with uniform emphasis across children affects some children more than others.

Practical Considerations and Conclusions

To restate the problem: Large-scale educational evaluations and government data systems often throw away useful information. This problem is not severe with intensive, small-scale studies; they have the time and resources to do thorough data analysis. But large-scale surveys, proposed "accountability" systems, and government information banks rely almost exclusively on average scores and average effects.

Given this situation and the continual need for policy decisions, there are three undesirable alternatives. First, one can forgo achievement data altogether, relying instead on less quantitative evaluative criteria. Second, one can choose to remain with average scores alone. Third, one may insist that evaluation cannot properly take place without a complete specification of educational objective functions for every level of government, every type of program and target population, all regions, every type of student, and, indeed, for every educational decisionmaker.

This paper has recommended a course of action different from all three. Although existing tests have shortcomings, some knowledge is better than none and therefore let us not abandon cognitive achievement measures. The mean is easy to use, but more knowledge is better than some, so we should go beyond simple averages. And although objective functions for evaluation are elegant, their practical application in education faces overwhelming obstacles.

The measures proposed here need further research before their exact properties are understood. Which exact statistics and which estimators to employ are open questions. As in the case of income distribution, there may be legitimate debate about which statistics are best. But also as in that case, the argument is that some such measures are better than none.

How should these statistics be used in the near term? Crude measures should be employed crudely. Continuous, cardinal uses of the statistics proposed would probably mislead more than they would help. A move away from pseudo-exactness is advisable. One might divide each percentile measure into five or so categories (say, the highest 20 percent of schools on each measure would receive a one, the second 20 percent a two, and so forth). (See also Dyer, 1972.) One might then envision a scheme like that shown in Table 4.

One should resist the temptation to concoct a grand measure, some weighted sum of all ten suggested statistics. Weighted sums assume mutual preferential independence, which does not hold for the proportion measures mathematically, and probably does not hold (given most reasonable objective functions) for any of the measures. Although complicated algorithms expressing conditional preferences are possible, it is best not to include these formally in any data system, accountability scheme, or large-scale evaluation. Let each decision-maker (and each citizen) be his own judge.

To propose the introduction of new measures without clear-cut objectives flies in the face of rationalist predispositions. But new measures even imperfect ones, can be the first step toward educational change. James March has suggested that most rethinking of objectives that does take place in organizations occurs precisely in a "backward" fashion--from changes in performance indicators to changes in goals and operations.

Using new statistics may shift discussions between educators and evaluators from questions of overall levels of performance to questions of equity, mobility, special programs, and the rest. One might imagine tables that show the tradeoffs among objectives that choices of different policies imply. The new statistics would not only more faithfully reflect the multiple and varied nature of educational objectives, they might also stimulate new concerns and create new incentives for action (or avoid some unwelcome old ones).

Looking for outliers in education and looking beyond a school's average score are two steps away from simple-minded evaluations. Unfortunately there are no well-developed methodologies or canned programs for doing either. There is much art, and much judgment, in discovering exceptional performers; in addition, there are normative questions involved in deciding which statistics to call the measures of school success. Computational ease and custom favor the use of means. Policy relevance favors going beyond them.

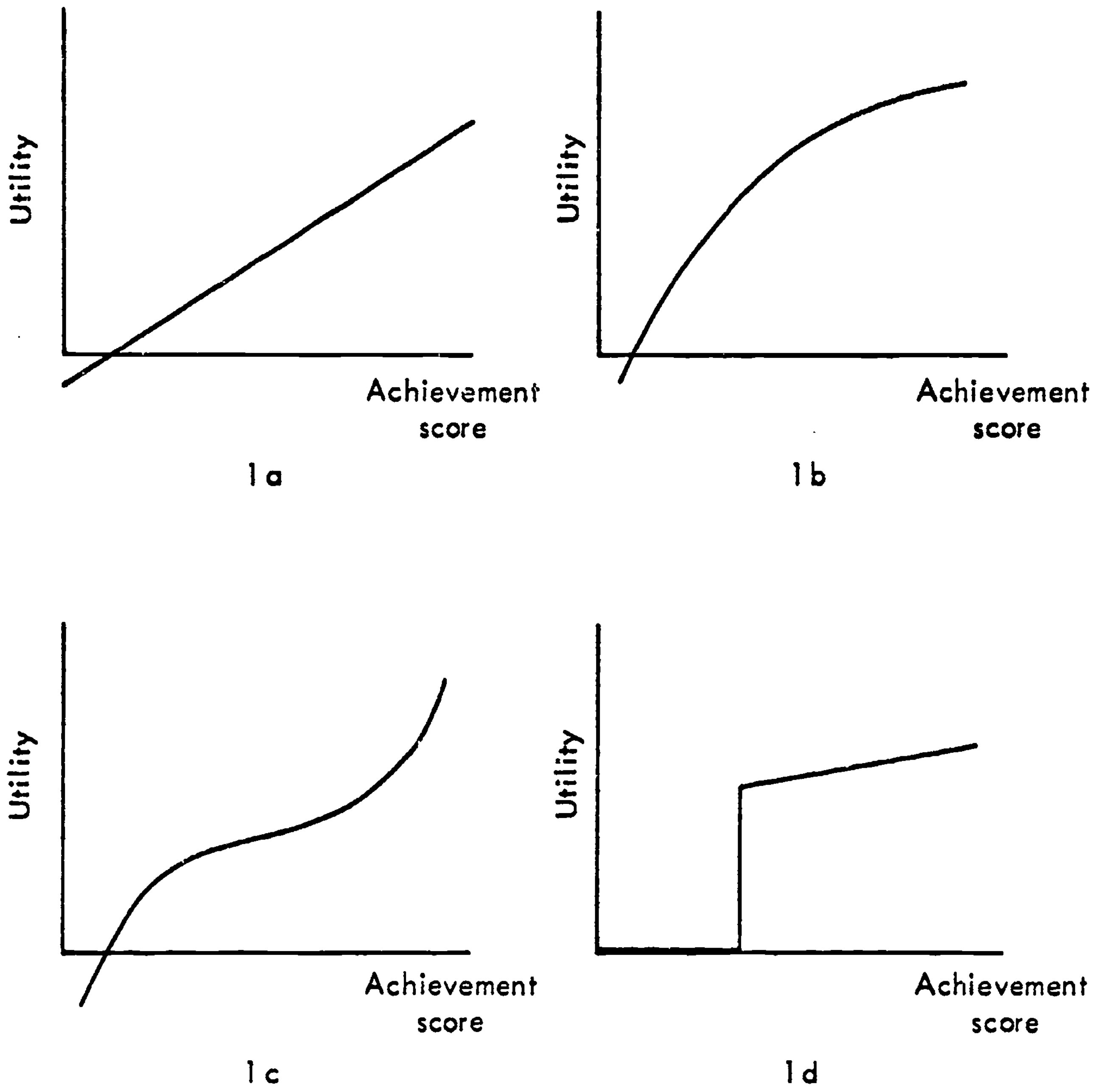


Fig. 1 — Some plausible shapes for utility functions for achievement

ERIC

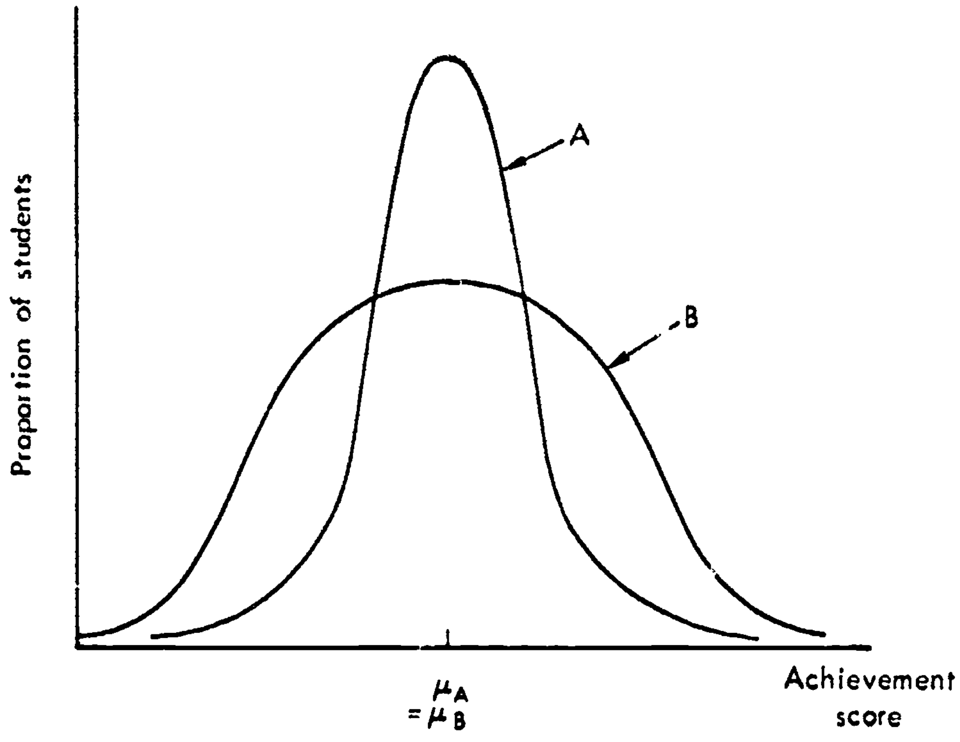


Fig. 2 — Schools with equal means and unequal spreads

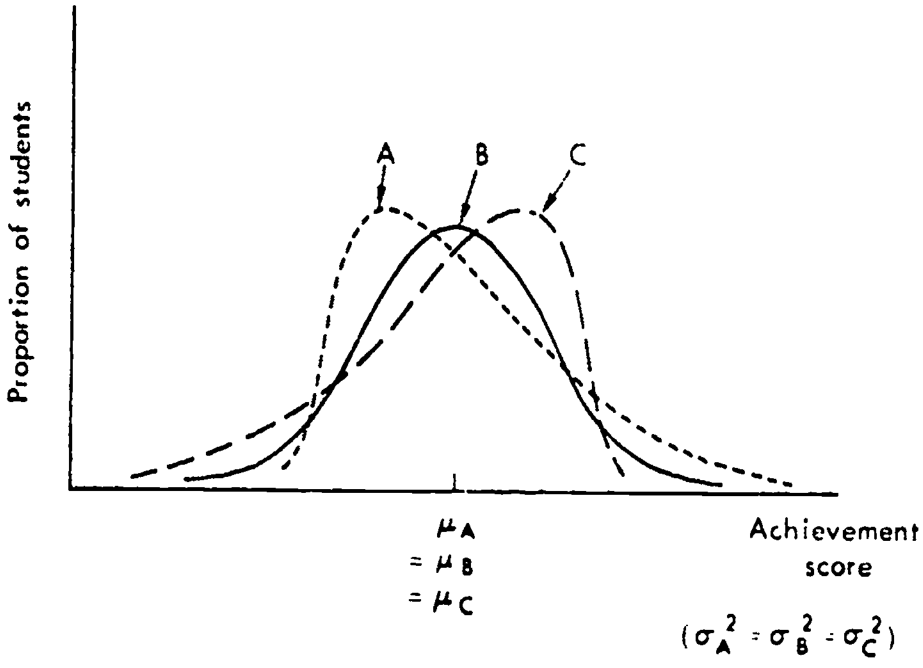


Fig. 3 — Schools with equal means and variances but unequal skewness

Table 1

MICHIGAN SCHOOL REGRESSION ON STANDARD DEVIATIONS OF TEST SCORES

Test	Equation	R^2	Std. Error	F	Number of Schools
R469-70	$\sigma = 17.8 + 0.14\ (SES\sigma) - 0.20(\mu) - 0.03\ (\%MIN) + 0.003(N)$ (72.5) (764.4) (458.8) (25.6)	0.26	1.03	206.9	2376
M469-70	$\sigma = 7.7 + 0.12\ (SES\sigma) + 0.003(\mu) - 0.01\ (\%MIN) + 0.004(N)$ (49.8) (0.2) (156.7) (35.1)	0.13	1.04	86.1	2376
R769-70	$\sigma = 9.7 + 0.17\ (SES\sigma) - 0.04(\mu) - 0.01\ (\%MIN) + 0.001(N)$ (35.0) (9.5) (38.0) (10.7)	0.11	0.88	25.5	870
M769-70	$\sigma = 8.1 + 0.19\ (SES\sigma) - 0.01(\mu) - 0.02\ (\%MIN) + 0.001(N)$ (40.9) (1.2) (125.4) (19.7)	0.25	0.92	71.5	870
R470-71	$\sigma = 19.9 + 0.08\ (SES\sigma) - 0.24(\mu) - 0.03\ (\%MIN) + 0.006(N)$ (28.6) (1147.4) (453.2) (79.4)	0.37	1.03	345.0	2401
M470-71	$\sigma = 8.5 + 0.10\ (SES\sigma) - 0.01(\mu) - 0.02\ (\%MIN) + 0.004(N)$ (41.8) (2.6) (188.7) (36.2)	0.32	1.07	67.3	2401
R770-71	$\sigma = 15.0 + 0.17\ (SES\sigma) - 0.14(\mu) - 0.02\ (\%MIN) + 0.001(N)$ (54.6) (140.8) (123.2) (14.4)	0.23	0.86	60.9	841
M770-71	$\sigma = 6.2 + 0.13\ (SES\sigma) + 0.03(\mu) - 0.02\ (\%MIN) + 0.001(N)$ (33.2) (7.0) (80.4) (30.3)	0.22	0.87	59.1	841

R469-70 stands for the reading scores for the fourth grades in 1969-70. The other symbols are interpreted similarly (M = mathematics). σ = test score standard deviation; $SES\sigma$ = socio-economic status standard deviation; μ = test score mean; %MIN = % pupils of minority races; N = number of students tested. Figures beneath the regression coefficients are the F-ratios ($=t^2$).

Table 2

MEANS AND STANDARD DEVIATIONS OF REGRESSOR AND RESPONSE VARIABLES

	μ	σ		μ	σ
R469-70μ	50.5	4.0	SES 469-70 σ	8.8	1.4
σ	8.9	1.2	SES 769-70 σ	8.6	1.2
N	62.8	34.7	SES 470-71 σ	8.8	1.4
M469-70μ	50.5	4.0	SES 770-71 σ	8.8	1.4
σ	9.0	1.1	%MIN 469-70	10.7	23.7
N	62.8	34.7	%MIN 769-70	10.5	22.6
R769-70μ	50.3	3.2	%MIN 470-71	10.1	22.8
σ	9.2	0.9	%MIN 770-71	9.4	21.1
N	172.7	130.2			
M769-70μ	50.4	3.8			
σ	9.0	1.1			
N	172.4	129.8			
R470-71μ	50.6	3.9			
σ	8.9	1.3			
N	63.5	34.4			
M470-71μ	50.6	4.2			
σ	8.9	1.1			
N	63.4	34.3			
R770-71μ	50.6	3.3			
σ	9.2	1.0			
N	182.5	133.9			
M770-71μ	50.6	3.9			
σ	9.0	1.0			
N	182.1	133.3			

Table 3

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

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	52	63	0	61	75	0	1852	1742
1	26	40	1	33	36	1	484	552
2	23	25	2	20	20	2	186	257
3	13	8	3	17		3	49	34
4	15		4	4 } 23	6	4	21	6
5	4		5	2				
6	2 } 25	3	6	0				
7	2							
8	2							

Chi-square = 167.0
Degrees of freedom = 4

Chi-square = 49.0
Degrees of freedom = 3

Chi-square = 74.3
Degrees of freedom = 4

All residuals were derived from a fit of the achievement score standard deviation against SES standard deviation, achievement score mean, percent minority enrollment, and number of students tested. All chi-square statistics are significant beyond the 0.005 level.

Table 4

EXAMPLE OF THE USE OF NEW ACHIEVEMENT STATISTICS

Educational Objective	Achievement Measure	School Number				
		101	102	103	104	etc.
General achievement level	Mean	2	5	3	3	
Achievement relative to student background	Residual mean	4	3	1	4	
Equality of achievement	Spread (perhaps σ)	1	3	2	4	
Equalizing effect of school	Actual minus expected spread	3	1	2	2	
Mobility afforded by school	Residual spread	2	4	2	5	
Effectiveness with exceptional children	Distortion (perhaps skewness)	1	3	2	5	
Effectiveness with over- and underachievers	Residual distortion	3	5	1	3	
Assuring children achievement skills at minimum level K	Proportion of students (A ≥ K)	2	4	2	4	
Assuring children do not under-achieve below level C	Proportion of students (R < C)	5	1	4	1	
Success with children above (below) background level S	Mean score of students above (below) S	3	3	1	5	

Numbers under schools refer to the following table:

Percentile	Category
80-100	1
60-80	2
40-60	3
20-40	4
0-20	5

Percentiles are computed for each statistic.

REFERENCES

Arditti, Fred D., "Risk and the Required Return on Equity," Journal of Finance, Vol. 22, No. 1, March 1967, pp. 19-36.

Averch, Harvey A., et al., How Effective Is Schooling? R-956-PCSF/RC, The Rand Corporation, March 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.

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

Craeger, John A., "Orthogonal and Nonorthogonal Methods for Partitioning Regression Variance," American Educational Research Journal, Vol. 8, No. 4, November 1971, pp. 671-676.

Cronbach, Lee J., and Lita Furby, "How We Should Measure 'Change'--Or Should We?" Psychological Bulletin, Vol. 74, No. 1, July 1970, pp. 68-90.

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.

Dyer, Henry S., Robert L. Linn, and Michael J. Patton, "A Comparison of Four Methods of Obtaining Discrepancy Measures Based on Observed and Predicted School System Means on Achievement Tests," American Educational Research Journal, Vol. 6, No. 4, November 1969, pp. 591-605.

Fisher, Irving N., and George R. Hall, "Risk and Corporate Rates of Return," Quarterly Journal of Economics, Vol. 83, No. 1, February 1969, pp. 79-92.

Friedman, Milton, and Leonard J. Savage, "The Utility Analysis of Choices Involving Risk," Journal of Political Economy, Vol. 56, No. 4, August 1948, pp. 279-304.

Guba, Egon D., "Development, Diffusion, and Evaluation," in Terry L. Eidell and Joanne M. Kitchel, eds., Knowledge Production and Utilization in Educational Administration, ERIC: ED 024 112, 1967.

Hicks, John R., Value and Capital, Second Edition, Oxford University Press, London, 1950.

Jencks, Christopher S., et al., Inequality, Basic Books, New York, 1972.

Klitgaard, Robert E., and George R. Hall, "Are There Unusually Effective Schools?" Journal of Human Resources, forthcoming.

Klitgaard, Robert E., and George R. Hall, *A Statistical Search for Unusually Effective Schools*, R-1210-CC/RC, The Rand Corporation, March 1973.

Kristol, Irving, "About Equality," *Commentary*, Vol. 54, No. 5, November 1972, pp. 41-47.

Lindquist, E. F., and A. N. Hieronymus, *Iowa Tests of Basic Skills: Manual for Administrative, 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, Washington, D.C., 1969.

Mosteller, Frederick, and Daniel P. Moynihan, eds., *On Equality of Educational Opportunity*, Random House (Vintage Books), New York, 1972.

Raiffa, Howard, *Decision Analysis*, Addison-Wesley, Reading, Mass., 1968.

Raiffa, Howard, *Preferences for Multi-Attributed Alternatives*, RM-5868-DOT/RC, The Rand Corporation, April 1969.

Raiffa, Howard, "Tradeoffs under Certainty," 1971 (unpublished).

Roche, J. G., *Investigation of Cost-Benefit and Decision-Analytic Techniques in Local Education Decisionmaking,* D.B.A. dissertation, Graduate School of Business Administration, Harvard University, 1971.

Tintner, Gerhard, "A Contribution to the Non-Static Theory of Choice," *Quarterly Journal of Economics*, Vol. 56, No. 2, February 1942, pp. 274-306.

Tukey, John W., *Exploratory Data Analysis* (limited preliminary edition, three volumes), Addison-Wesley, Reading, Mass., 1970.

Vaisey, John, *The Economics of Education*, The Free Press, New York, 1962.

Von Neumann, John, and Oskar Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, 1944.

Yarnold, James K., "The Minimum Expectation in X^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.