DOCUMENT RESUME

ED 102 677 EA 006 798

AUTHOR	Klitgaard, Robert E.
TITLE	Improving Educational Evaluation in a Political Setting.
INSTITUTION	Rand Corp., Santa Monica, Calif.
REPORT NO	P-5327
PUB DATE	Dec 74
NOTE	39p.
AVAILABLE FROM	Rand Corporation, 1700 Main Street, Santa Monica, California 90406 (Order No. P-5327, $3.00)
EDRS PRICE	MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS	Academic Achievement; Bibliographies; Case Studies (Education); Data Analysis; Data Collection; *Educational Accountability; *Educational Policy; *Evaluation Criteria; Information Processing; Information Utilization; *Political Issues; *State School District Relationship; Statistical Analysis
IDENTIFIERS	*Fulano

ABSTRACT

Evaluations that use imperfect information run into both analytical and political problems. Educational accountability systems based on achievement scores are an instance. Such systems frequently turn out to be irrelevant to policy decisions, resisted by educational interest groups that fear unflattering comparisons and the misuse of results, and infeasible. This paper suggests several ways to analyze limited information that both improve the decisionmaking relevance of evaluation and decrease the adverse political consequences. The context for discussing these issues is a case study of an early stage in the design of a State accountability system for public schools. The paper discusses a series of theoretical and practical problems in assessing the effects of school policies with imperfect achievement data and limited controls for nonschool factors that influence scores. Problems of the collection, analysis, and presentation of data are explored in the case study. Hypothetical regression equations and tables of policy-relevant results are provided as illustrations. (Author/WM)

IMPROVING EDUCATIONAL EVALUATION IN A POLITICAL SETTING

Robert E. Klitgaard

December 1974

P-5327

$3.10

The Rand Paper Series

The Rand Corporation
Santa Monica, California 90406

SUMMARY

Evaluations that use imperfect information run into both analytical and political problems. Educational accountability systems based on achievement scores are an instance. Such systems frequently turn out to be irrelevant to policy decisions, resisted by educational interest groups that fear unflattering comparisons and the misuse of results, and infeasible given faulty data and limited time and money.

Accountability systems in state governments are increasingly widespread, and state governments seem to be more and more important as educational policymakers. In the state I rename Fulano, accountability ran into political opposition and feasibility constraints which are instructive to examine. Policymakers consistently and mistakenly saw important statistical issues as merely "technical" questions. Actually, these statistical issues were at the heart of more relevance, less resistance, and greater feasibility.

The paper discusses a series of theoretical and practical problems of assessing the effects of school policies with imperfect achievement data and limited controls for nonschool factors that influence scores. Problems of the collection, analysis, and presentation of data are explored in the case of Fulano. Hypothetical regression equations and tables of policy-relevant results are provided as illustrations. Practitioners may find the discussion useful in designing an accountability system, although many of the ideas presented have not yet been tested in practice.

The broad recommendations of the paper are listed below, although they are probably not too illuminating without the specific details of their application that are provided in the body of the paper. Accountability systems, and evaluations in many other policy areas besides education, could be improved by:

(1) Employing multiple measures of outcomes, which increase policy usefulness while tending to reduce the misuses and opposition that accompany one-dimensional listings of schools and districts.

(2) Comparing policy choices, not just schools and districts, which relate variables that state decisionmakers can affect to outcomes and reduce the reluctance of local officials to let their data be used.

(3) Searching for unusually effective schools, which are relevant to policy and also are likely to provide stories of success that can lessen the chance of misleading generalizations about the failure of education.

(4) Stressing, in both words and the way the numbers are presented, the uncertainty and inexactness of the evaluation's findings.

CONTENTS

IMPROVING EDUCATIONAL EVALUATION IN A POLITICAL SETTING*

Robert E. Klitgaard

The Rand Corporation

1. Introduction

Public sector evaluations offer the hope of improved policies. More knowledge, say the optimist, the scholar, and the Platonist, leads to better behavior--fewer mistakes, faster diffusion of successful innovations, more thorough planning. But there is also a pessimistic view of the use of information in the public sector. Critics contend that, in a political setting, evaluations may only be the first step toward adverse incentives, incorrect decisions, and harmful in-fighting between the various parties who are potentially affected by the changes or misuses that evaluations may portend; or that evaluations, even when theoretically useful to decisionmakers, may simply be ignored.

Moreover, when, as is usual, one has only partial and inexact information on which to base such evaluations, the optimist must be more guarded and the pessimist can scowl more fiercely. Indeed, in the case of the evaluation of public schools, where achievement scores constitute one of the only widely available sources of information, it is difficult to be optimistic about the usefulness of many current evaluations for decisionmaking. Some are mere lists of mean scores by school; or, when regression techniques are used to control for school and nonschool factors, only the average effects of policies on school average scores are considered. And it is easy to be pessimistic about the possible adverse results of such information: mistaken generalizations about the ineffectiveness of schooling, potentially misleading and politically explosive comparisons among schools and districts, hostility from teachers and administrators.

---

*I am indebted to Will Fairley, Milbrey McLaughlin, Richard Light, David Mundel, and Richard Zeckhauser for their help on earlier drafts. Numerous state officials also deserve thanks, but they must remain unnamed.

Seldom do large-scale evaluations of public schools employ available information wisely and with an eye to policy relevance. Seldom, too, do governmental accountability systems consider how the prospects for misuse and misunderstandings of evaluations should affect the way they do their evaluating. This paper suggests several ways to analyze limited information that both improve the decision-making relevance of evaluation and decrease the adverse political consequences.

The context for discussing these issues will be a case study of an early stage in the design of a state accountability system for public schools. To protect confidences, I will call this state Fulano (after "fulano de tal," the Spanish analogue to "John Doe"). This paper will briefly review the history of state accountability systems both nationwide and in Fulano. It will then discuss how the seemingly "technical" issues of collecting, analyzing, and presenting achievement score data interrelate with the overriding political and organizational problems of a statewide evaluation, as well as with the problem of making the information collected useful to policymakers yet not prone to misuse and misunderstanding. I will try to show how certain improvements in the use of limited information could be implemented in a concrete setting. The story in Fulano is not yet over; but even a partial account of an early stage of the accountability system's design may be instructive, both as regards the difficulties of educational evaluation in general and the usefulness of going beyond mean achievement scores in particular.

2. Accountability Systems in Public Education

The idea of making local school authorities accountable to parents and to higher levels of educational governance is by no means new, but only in the late 1960's did the idea gain widespread acceptance in the United States.[1] Many forces contributed to the upsurge. For one thing,

---

[1] In 1971 Harold Spears listed the major educational policy issue for each year back to 1956. Accountability was his key issue for 1970 (cited in Bruno, 1972, p. 19). For a general history of state accountability systems, see Dyer and Rosenthal (1971).

educational expenditures were rising at an alarming rate,[1] and in the aftermath of the Coleman report, many taxpayers and politicians were skeptical about the effects of the increases. Federal expenditures for public schools had jumped dramatically in the 1960's,[2] and so, as a result, had federal interest in evaluating local education. Local interest in accountability was also high. New programs that encouraged citizen participation in school affairs often led, as one of their first results, to efforts at developing accountability procedures[3]--not surprisingly, since educational costs usually were responsible for over half the amount of local taxes.

"The basic theory of accountability," wrote Leon Lessinger, considered one of the key forces in the federal government responsible for the sudden growth of accountability, "is that school personnel have an inescapable responsibility to account for the accomplishment of the students entrusted to them in terms of specific performance objectives they--the educators--have publicly established as a condition for the receipt of resources" (1972, p. 231).

This "basic theory," however, was consistent with a number of forms of accountability: performance contracting, voucher systems, merit salaries, and systematic evaluations of local educational needs, objectives, outcomes, and costs. The latter category of systematic evaluations usually was performed by state governments, but local and federal levels were also frequently involved in their own accountability efforts.

---

[1]From 1950 to 1970 school enrollment went up slightly more than 80%; school revenues in price-adjusted dollars had increased some 350% over the same period.

[2]During the decade the Office of Education grew in budget from less than $500 million to more than $4 billion; in programs, from about 15 to approximately 100.

[3]In a thirteen-city study Cunningham and Nystrand found that "one of the most important purposes of new efforts by citizens to participate in school affairs was an accountability drive" (cited in Cunningham, 1972, p. 80).

Evaluations could be done of teachers, principals, school districts, school boards, programs, or even states.[1]

The efforts of state governments to assess schools and districts are of particular interest for several reasons. First, such efforts are widespread. By early 1972 thirty states had implemented some such accountability system, and all 50 states "reported assessment activities either as operational, in a development process or in a planning stage" (Educational Testing Service, 1973, p. 1).

Second, the state government has a central role in American educational policy. "State legislators, as a population, and especially the committee structures (education and appropriations)...comprise the most powerful decision group in education" (Cunningham, 1972, p. 82). As Wirt and Kirst point out, most state constitutions deem local education as a responsibility of the state; and state-level decisions include a large role in setting the level of funding and its allocation,[2] deciding the scope of programs, setting minimum standards, designing courses of study, and training and certifying teachers. They need accountability information for all these decisions. "Although popular folklore conceives of schools as locally controlled, the state has taken a hand in them for many decades. Indeed, by 1970...it is highly questionable how much local control is a reality" (Wirt and Kirst, 1972, p. 111).

Third, although there is considerable diversity among states in the degree of power of different bodies concerned with education, decisions

---

[1] For a comprehensive account of the various forms of and protagonists in different accountability schemes, see Sciara and Jantz (1972).

[2] "The governors and legislators, however, maintain control of state financial-aid legislation...In most if not all states, education is the largest single state budget item, and politicians, of course, know that the electorate responds to tax increases...The weight of such monies gives much power to those who dispose it, and in the states these are the governor and legislature" (Wirt and Kirst, 1972, pp. 114-5). States also have an increasingly important role in allocating funds from federal programs.

are usually made by a coalition of interested parties.[1] Seldom does one body have sole power to decide. "A recurring theme [in case studies of decision-making in the state] is the lack of decisive political influence over financial aid decisions of state government by any single educational organization or group" (Kirst, 1970, p. 216).

These three considerations mean that the information provided by accountability systems potentially has important consequences. Data for evaluation are in greater demand and are becoming more plentiful. They have importance for decisions by state officials about educational expenditures. And, in a context of increasingly unstable[2] political coalitions at the state level, the new inflows of information are bound to be perceived, and used, as means of reallocating power among the contending parties.

Typical Problems

There have been, I think, three types of problems common to state accountability efforts, especially to those which attempt to gather data about costs and student achievement for policy purposes:

(1) Irrelevance to policy. Even in states where data are obtained expressly to help legislators and administrators make better policy choices, accountability reports are often irrelevant. Usually a large statistical compendium is produced that describes the average levels of performance (on one or several achievement tests) of each school and district. Frequently the descriptions include summary statistics of some school resources, such as pupil/staff ratios, teacher salaries, and so forth. Occasionally, as in California and New York, each school or district is given an adjusted achievement score mean, after controlling for some nonschool factors by means of multiple regression or some sort of stratification. These descriptions are useful as descriptions. (For

---

[1]See, for example, Bailey et al. (1970) and Masters et al.(1964). The parties involved in the coalitions include the governor, the legislature, the State School Board, the chief state school officer, teachers' groups, local school boards, and superintendents' groups, as well as parent-teacher associations, taxpayers' groups, and parochial school bodies.

[2]Kirst (1970), pp. 216-219.

example, they may, despite the forbidding tables that are customary in such reports, help parents as indicators of where their chiildrens' schools stand. As such they can be an important first step toward local pressures for improvement.) But there is almost never any indication of how policy variables that state officials control affect student outcomes. Nor is there any attempt to locate systematically any policies or schools that seem unusually successful. As a consequence, policymakers and parents are presented with a mass of print-outs that are hard to comprehend and difficult to link to policy choices.

(2) Political opposition. Statewide assessment threatens local powers. Teachers fear that they may be held accountable for outcomes over which they have little control. Local educational institutions--school boards, superintendents, even colleges and universities that train teachers--worry that state evaluation may be the first step toward state usurpation. They also are often reluctant to have crude mean achievement scores published, feeling that such scores are imperfect, partial measures that will inevitably be seized by newspapers and local citizens as precise, complete evaluations. They see little usefulness and much possible misuse in the mere description of resources and scores.

(3) Partial and inexact information. Complete information on pupils' progress along the wide spectrum of educational objectives is impossible to obtain at present. "Accountability" must therefore proceed with partial information. Furthermore, even the limited information that is gathered is inexact. Sources of error that occur merely in estimating a school's mean achievement score include testing error, different and small sample sizes, and incomplete data collection. If one attempts to adjust a school's score to take account of the students' socioeconomic backgrounds and other nonschool factors, further errors creep in;[1] and most states do not possess extensive information on the

---

[1]For example: 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.

non-school factors that determine achievement. These imperfections are not necessarily fatal; an evaluation can be useful despite them. However, they can become extremely important when the results are presented or interpreted as a complete evaluation of a set of schools, or when the margin of error is not clearly indicated in the way the data are reported: both of which have frequently occurred in practice.

Minimizing these three problems of state accountability systems in the case of Fulano is the topic of the next sections.

3. History of Accountability in Fulano[1]

Accountability arrived relatively late to Fulano. One can hypothesize that this tardy start was due to a legislature that met only three months a year and had little educational expertise on its staffs, a fairly weak state department of education (FSDE), a strongly entrenched local educational structure with lobbying power, and a strong teachers' lobby. In 1970-71, FSDE initiated a task force on evaluation that considered, among other things, a competency-based teacher certification policy. Nothing much but dissension arose from these considerations (in the words of one Fulanese official, "a pooling of ignorance and frustration"). But by 1971, taxpayer pressure had, according to legislators, risen to the point where accountability became a key item on the General Assembly's agenda. After much work and discussion, a teacher accountability bill was presented for a vote, and, owing to the vigorous opposition of the State Teacher's Association lobby, it was defeated. As a result no bill was passed in 1971.

When it convened in 1972, however, the legislature was determined to enact some sort of accountability law. After considerable lobbying and debating, the General Assembly reached agreement on a compromise scheme: no teacher evaluation and no mandated testing (to please the lobby), but a definite (if open-ended) plan for state assessment of schools and districts. "The purposes of this Act," declared the legislators in the bill that finally passed,

---

[1]This history is based on interviews with officials in Fulano in November 1973 and January 1974, and published and private documents provided to me at that time.

> are to provide for the establishment of educational accountability in the public education system of (Fulano), to assure that educational programs operated in the public schools of (Fulano) lead to the attainment of established objectives for education, to provide information for accurate analysis of the costs associated with public education programs, and to provide information for an analysis of the differential effectiveness of instructional programs. (emphasis added).

When the assembly recessed at the end of spring, these three general goals--the assessment of objectives, costs, and differential effectiveness--were turned over to the State School Board and the FSDE for detailed elaboration by the following winter, when the Assembly would convene again.

FSDE was the subject of a good bit of pressure as to how it would design the specifics. The chief state school officer (CSSO) indicated, to the pleasure of the pressurers, that statewide testing would not be required. In the meantime, a blue-ribbon task force, dominated by FDSE's coordinator of planning, created, by December 1972, a very detailed 130-page document that allowed only for evaluation on the basis of local objectives and did not recommend statewide, norm-referenced testing. But this was still too much evaluation for local interests. In December, the 24 district superintendents met--not in the state capital--and unaminously adopted a resolution condemning the proposed model system. The scheme was also rejected by the State School Board, which had the official power to decide. Some Board members sided with the superintendents that too much state intervention was being attempted, while others, including some of the most powerful members, claimed that not enough in the way of objective measures were included.[1]

---

[1]The "official" explanation for the failure of the plan of the blue-ribbon task force is provided in the accountability Final Report 1973 (pp. 1, 11):

> Subsequent to the presentation of the accountability model implementation guidelines to the LEA superintendents and to the State Board of Education, the guidelines were not approved. The negative reaction to the guidelines was based essentially on the general opinion that they were too lengthy and that unnecessary verbiage

With eight months' work rejected, accountability--and FSDE--were clearly in trouble. The legislature convened and was irate at the lack of action. The failure of FSDE to make progress confirmed an increasingly popular view that the educational system was rigid, wasteful, and expensive. There was talk of reorganizing FSDE, the only major bureaucracy not to have been affected by sweeping state reforms during the last five years. The sabre-rattling was severe enough to prompt a speedy response from FSDE and the State School Board. A new accountability chief was brought in, and within a month the Board had approved a plan that would (1) involve statewide testing and (2) use local people whenever possible in designing the system, to reduce local resistance.

An Advisory Council on Accountability (ACC) was created, composed of 23 members who were selected from non-educators as well as from parents' groups, teachers' organizations, superintendents, legislators, and university professors. The ACC would report its recommendations to the State Board. In addition, a group of Local Coordinators for Accountability (LCA) was created, appointed by superintendents and meeting in four regional groups. The new FSDE accountability chief was a leading participant in both the ACC and LCA.

The big issue was still statewide testing: Would it occur? And if there were to be testing, which tests, to what extent, across which bodies of students? Some locals still hoped to block statewide testing, and they applied pressure through the ACC and the LCA. Finally, after lengthy ACC debates, in June 1973 the group voted 17-6 in favor of statewide testing using the Iowa Test of Basic Skills (ITBS). A deciding factor was allegedly a threat by four legislators to draft a much more specific accountability bill if no statewide testing occurred.

The rest of 1973 was spent deciding how many tests were to be used, which grades, whether some districts would be allowed to sample students, and if special students should be included. Many of these issues were still unresolved by January 1974. After the principle of testing was approved, both the ACC and LCA's found themselves locked into discussing

---

obscured the essential contents...(L)ocal constituencies of the (Fulanese) State Department of Education did not respond favorably to a comprehensive model for Accountability. That does not mean, however, that the model that was developed is inappropriate. Perhaps it was only overwhelming in its comprehensiveness.

technical issues and making recommendations with little impact.[1] The focus of decisionmaking shifted to MSDE, which busied itself with the logistics of statewide testing, by no means a trivial task.

But with all the concern over the existence and the logistics of statewide testing, little attention was paid to how this information and other data would be used in an analysis of school and district success. There were efforts by the LCA to draft local objectives along which success could be gauged, but there was little progress. The ACC drafted a set of state educational goals, resounding in generalities and existentialist rhetoric, but useful only in its negative implication that the statewide testing would measure only part of the state's goals. There was also discussion on the possibilities of using regression analysis to control for nonschool factors that differed among schools, but again little was decided. Many questions of data collection, analysis, and presentation were still unanswered in late 1973, and the first report was due only a little over a year later.

4. Decisions and Data in January 1974

An advisory group of four outside consultants, of which I was one, was summoned in November 1973 in an effort to decide how the achievement data should be analyzed. FSDE saw this problem of analysis as predominately a technical one: aggregating the information, perhaps "standardizing" school scores according to socioeconomic or other nonschool factors, and writing a technical report.

Little attention had been paid, however, to the way these "technical questions" of collection, analysis and presentation interfaced with the three broad objectives the Fulanese accountability system really had.

(1) Policy relevance. The first goal was to provide useful information to policy makers. What decisions need to be made? What information would make these decisions more cost-effective, and how much more?

---

[1] In two of the fall 1973 ACC meetings, the Chairperson declared that the group needed a clearer definition of its purposes from the CSSO. After May 1973, according to the accountability chief, the LCA's had ceased in their "advisory role" and were used primarily as conduits of decisions made by FSDE.

How much does the information cost? Such questions are part of a decision analysis model, where the key idea is to collect information only to the point where its expected usefulness exceeds its cost.

Thus described, this first objective may seem overly academic. But emphasizing the decisionmaking value and the costs of accountability information is an important antidote to the real academic menace--to view accountability as a research project, a journal article, a collection of interesting statistics. Too often evaluation systems end up providing reams of information that do not link the policymakers' (many) objectives with their possible choices of action. How accountability information can eventually be used should always be a primary question, and the answer should largely determine the data to be collected, the way they are analyzed, and how results are presented. For FSDE, however, the question was largely seen as one of simply describing the test results.

(2) The political aspects of evaluation. There is a second goal of evaluation: to prevent the misuse or misunderstanding of the results. Often these adverse results cannot be avoided altogether. Often, though, there is no necessary black-and-white choice between "don't include that information" and "let the facts fall as they may." Hostility-provoking data can be reported cautiously and ingeniously; misuse can sometimes be avoided by the judicious choice of tables and words.

(3) Existence. In light of formidable problems of logistics and implementation, an important goal was "merely" to have statewide testing for the first time in many years, to produce a report based on test results, and not to discredit accountability efforts of the future with a failure (or an antagonizing success) the first time through. A major goal of an evaluation was to exist--or in this case, to come into being. Like nations suffering the attacks of hostile aliens, many accountability efforts see their own survival as the major task. Evaluations can become a version of publish or perish, where the fact that a volume is produced matters more than what it says. When the survival of the accountability system as well as the survival of particular personalities are at stake, existence is no trivial objective. Yet existence was not a function solely of technical feasibility; local resistance was still a

major consideration; and the choice of methods of analysis could have a sizable influence on that resistance.

I urged that these three objectives be taken as the basepoint from which to evaluate alternative strategies of data collection, analysis, and presentation. To treat the problem as simply a technical one was to miss much that mattered. In January 1974, I returned for a more extensive visit to try to translate those generalities into specific steps. In what follows, I will not report the entire range of recommendations, nor touch on all the questions that an accountability system must address, but will emphasize the implementation of various methods for the improvement of statistical analysis. I will divide my remarks into three sections: data collection, data analysis, and presenting the results.

4. Data Collection

The Iowa Test of Basic Skills, the chosen measurement instrument for the legislature's mandate to "survey the current status of student achievement in reading, language, mathematics, and other areas," is superbly normed and highly regarded among testers, but "in every subject matter area ITBS measures only a very small portion of one or two, if any, of the goals adopted" (Advisory Council on Accountability, "Design for Accountability," p. 3).

If an evaluation is necessarily incomplete, it will be natural to expect resistance from those who fear misuse of partial measures. Test scores are clearly only part of what an evaluation should be; and if they are the only part that our current measurement situation will allow us to have, we must make plain in words and in the way we use the numbers that we understand the limitations.

Even if an evaluator only has achievement scores as measures of success, he can still look at statistics of the school and district distributions of scores that have intuitive links to broad objectives.[1] And he can control for nonschool factors that affect achievement scores but which are not uniformly distributed throughout the state. Reliance

[1] See Klitgaard (1974).

on mean scores alone is a step toward the "one-number-per-school" sort of evaluation that will not only mislead but will throw opponents of accountability into a snit. One major recommendation, therefore, was to use multiple measures of school outcomes in a regression analysis framework.

What data would be necessary for such an analysis, and how could they be gathered?

Fulano had already decided to administer eight of the Iowa Tests of Basic Skills as achievement measures and the non-verbal Cognitive Abilities Test as a proxy for (non-school-related) intelligence. This battery would be given to all third, fifth, and seventh grade students (except for one district, which wanted to sample students from each grade).[1] Eighteen districts had opted for scoring by an outside organization, which would provide means, standard deviations, and certain percentile values (10, 25, 50, 75, 90) for both achievement and intelligence tests by grade and school, as well as individual data for use by school counselors. The other six districts preferred to do the scoring and the reporting themselves. The problem was that they did not wish to report much to the state, and FSDE finally asked only for means and standard deviations by grade and school, but not the percentile data.

The six districts were already complaining about programming burdens. Since none already used the eight tests of the 1971 ITBS Form 5, new scoring programs would have to be developed. Some districts did not have an interest in standard deviations; one preferred the median to the mean; therefore, all felt encroached upon (or pretended to) by even such simple reporting tasks. There was the fear in FSDE that come May or June 1974--after the testing had been completed--even districts that were then confident of their ability to get tests scored and reported by the late June deadline would realize that the task was difficult, or would find their computers bottled up with end-of-the-

[1] Amusingly, the CSSO maintained that he had kept his promise that there would be no state-wide testing, because one district would only sample its students.

school budget matters, or would suffer other problems. And if the data were not submitted to FSDE on time, summer vacation might mean an additional three-month delay (at least). Such an event might delay the whole report.

In this context, it might seem overly ambitious to request additional statistics, despite their analytical usefulness. One would like to increase the data requested from the six districts to include at least the percentiles (in order to capture objectives like "success with slow students," "success with bright students," "equalizing effect of the school", and others). But one was also wary that more data requests might create such hostility or overload (or both) on the part of local officials as to endanger the timing, or even the existence, of the first statewide report. It was tempting to say "Better 'something' than 'more' if 'more' means 'nothing,'" and stick to simple averages.

However:

(a) Policy actions could be taken by FSDE to alleviate the problems of formating and scoring tests at the local level.

(b) The individual-level tapes produced by each district--which would have to be produced in any event by districts to calculate the means and standard deviations--could be duplicated and sent to the state for the state to produce the requested and the additional statistics (see Klitgaard, 1974).

Such actions might alleviate risks of delay and bureaucratic resistance as well as enabling data comparable to the other 18 districts to be obtained. Neither course was being considered by FSDE.[1] I recommended that both be investigated at the first opportunity.

---

[1]The state might have covered itself by requesting both means and standard deviations and the tapes. That way, if there were a foul-up at either end, at least there would be something to analyze. The scoring problem was a serious logistics problem that the state should immediately attempt to treat. It could provide the cause or the rationalization for a district's non-participation.

In what follows I assume that the state is able to obtain only four statistics--the mean, standard deviation, percentage of students below the national tenth percentile, and the percentage of students above the national ninetieth percentile.

Other Information

Apart from test scores, other data should be gathered. A school's achievement scores are a function not just of the "intelligence" of its students as measured by the Cognitive Abilities Test. They are also presumably a function of the level and allocation of resources among different curricula, facilities, and policies, and of the backgrounds of its students. If one is interested in the impact of education on scores--and specifically that of different policy choices--one must control for these additional variables as well. FSDE had considered these factors in a general way, but it had not moved to collect the necessary data on school, district, and socioeconomic variables.

School variables. Surprisingly, FSDE was not planning to analyze the impact of differential school policies, despite the legislation's plea for such information. Which such data should and could be collected? Unfortunately, there is no convincing model of the variables that should be included to capture a school's entire contribution to student achievement gains. Furthermore, the state had limited school-based information, confined to characteristics of the average teacher and average administrator (salary, educational level, experience), student-staff ratio, and age of the building. Of course, the impact of these few variables on achievement scores will not summarize the impact of the school, but this limited information is of interest and should be included. The point is that legislators and state officials made decisions about resource allocation across various policies, such as teacher salaries and tenure, size of classrooms, the provision of school facilities, curricula, and administrative variables. Many interest groups argued strongly to the legislature that certain levels of provision of these policies were essential for successful instruction, that better-paid teachers led to better learning, that small classes enabled students to master basic cognitive skills better than large classes, and so forth. Others, however, argued that these differences had been shown in other places and times not to be important. Debates could not be resolved by citing the results of studies that took place elsewhere, for the advocates of more education would argue that those results were not necessarily true in

Fulano. Thus, there was a need to assess, in Fulano over the recent past, the effect on student learning of basic skills that various policy choices would have. Even if an accountability system would not gauge the entire contribution of a school and even though the assessment of certain variables would have to take account of the partial and imprecise nature of achievement scores, it would be useful to state policymakers to know the relationship between those variables and those outputs in a general fashion.

The state was not planning to collect or analyze this information at the school level, a decision I worked hard to try to reverse.

District variables. The state's data system already had a ready print-out of district means on the above-mentioned variables. Since one could at little cost also get them at the school level, one should. But since the effectiveness of district resources and policies was also of interest, those variables should also be included at the district level. Furthermore, district-level information was available on per pupil total, instructional, administrative, and personnel services costs, none of which could be broken out by school. These proxies should also prove interesting to relate to achievement scores, even though they of course would capture only part of the district's "true" impact and though they represent costs often pursued for non-achievement objectives.

Socioeconomic and background variables. The desire to collect non-school background variables in an analysis of school and policy success in imparting basic cognitive skills stems from the fact that the former greatly affect test scores. A crude model of how scores are determined is

$$\text{Achievement} = f(\text{School, Home Background, Genetic Factors,...})$$

Estimates of the school effect (or the effects of various policies schools use) that do not control for differences in home background and genetic factors will be biased.

While no one denies such a general formulation, problems arise in trying to define and to measure "home background" and "genetic factors." No one "knows" what the measurable dimensions of "home background" really

are for these purposes: usually the concept is defined via certain measureable proxy variables that have predictive power and are congruent with the sociologist's intuition that one's past environment determines at least part of one's present success. One such proxy is socioeconomic status (SES), and economists frequently use proxies for that.

In the case of Fulano, where little information in the way of "home background" or "genetic factors" existed statewide, what biases in the estimates of school and school policy effects would be introduced by using various proxies for those variables?[1] There are several general ways of dealing with this question.

The first is to assert that one is not trying to estimate the true effect of "home background," "genetic factors," or "school effects," but instead to propose, quite empirically, certain controls that enable one to better assess the effect of different schools and school variables. On this pragmatic view, one announces findings after controlling for certain limited nonschool variables X, Y, Z..., which turn out to correlate with other factors of interest in certain specific ways (for example, achievement scores and school variables). There is no pretense of having controlled for some larger variables (or concepts) like home background or genetic factors. One's findings are purely empirical descriptions of the relationship of various measurements; no theory is tested nor are parameters of a preexisting formal model estimated.[2]

A second point of view examines the robustness of certain possible policy-related findings to the specification of the control variables. For example, if one only possesses limited data on certain aspects of socioeconomic status (which is itself only one proxy for "home background"), how good are those aspects as proxies? How sensitive are estimates of school effects to the inclusion or exclusion of certain proxies?

To answer this, first let us examine the data available in Fulano. Somewhat outdated 1970 census data existed by districts, but not by

---

[1] A concise theoretical treatment of the biases involved in omitting a variable or using a proxy variable is available in Rao and Miller (1971), pp. 32-4, 82-8, and 115.

[2] See Coleman et al., (1966); Coleman (1970); Jencks et al., (1972); Smith (1972). This position has also been criticized: see, for example, Cain and Watts (1970) and Porter and McDaniels (1974).

schools. The so-called Urbanetics data provided two-year-old SES proxies by school area, but because of bussing and open enrollment policies, in certain districts data for a school's area no longer corresponded with the SES of its actual students. The only proxies that were computed from a school's students were its racial composition and the numbers of free lunches applied for and received. Race was a touchy subject; the Advisory Council had already voted twice against any mention of it in the report. Free lunches are not a good indicator of disadvantagedness, because not all children eligible actually apply.

Census data by school are used often as a proxy for SES and predict school mean achievement scores fairly well (R^2 = 0.3 - 0.4), but in this case the census data are (a) fairly old and (b) not accurate for schools with bussing. Racial information is also a good predictor of achievement, and it correlates positively with SES (Jencks et al., 1972, pp. 81-3). Free lunches do not correlate highly with the six-variable SES index estimated by Coleman, et al., (1966)--the correlation is about 0.35 (Jencks, in Mosteller and Moynihan, 1972, p. 94). But even the detailed SES measure devised by Coleman et al., does not capture much of true influence of "home background." The latter, as estimated, for example, in studies of identical twins reared together and apart, explains about 35% of the variance in achievement test scores; economic status explains only about 6% (Jencks et al., 1972, p. 109 and Appendix A). So, in terms of the percentage of variation in test scores that these various proxies explain, we conclude (1) they are not perfect proxies for home background and (2) they nonetheless do correlate positively with home background, and they do explain variation in achievement scores for which schools can hardly be held responsible.[1]

But what about the bias introduced by using proxies? How robust are estimates of school effects to the inclusion of different SES measures? The desired estimates can be conveniently grouped into three classes: (a) the effect of different facilities and curricula on achievement scores; (b) the effect of various teacher characteristics; and (c) rankings of school scores.

---

[1]See Coleman et al., (1966) for a discussion.

(a) Estimates of the effects of facilities and curricula. It turns out that these estimates are extremely robust with respect to the inclusion or exclusion of various SES proxies. In terms of the variation of mean achievement scores explained, Smith, working with northern schools from the EEOS, found that "the proportion of variance uniquely accounted for by the Facilities and Curriculum factor decreases slightly when it is not first controlled for individual home background conditions" (in Mosteller and Moynihan, 1972, p. 241). More importantly, the standardized regression coefficients (Beta weights) do not significantly change with the inclusion of more or fewer SES controls:

> The extraordinary thing about these final four columns [in a table comparing the estimates of the effects of six different school policy variables under four different SES controls] is the similarity in the magnitudes of the Beta's for each of the variables. With only a few exceptions the Beta's in the equation with all of the [SES proxy] variables entered are as large as the Beta's in equations with less controlled conditions (Smith in Mosteller and Moynihan, 1972, pp. 244-6).

(b) Estimates of the effects of teacher characteristics. The teacher characteristics available in Fulano include a proxy for experience (number of years), average teacher pay, and student/teacher ratio. The estimate of the effect of teacher experience (as measured by Beta weights) is very robust with regard to the inclusion or exclusion of SES information (Smith in Mosteller and Moynihan, 1972, p. 245). The other variables have not been examined. However, in light of the general findings by Coleman et al. (1966) and Jencks (in Mosteller and Moynihan, 1972, pp. 73ff) that there is virtually no relation between either the racial composition of a school or its socioeconomic level and the level of a large number of school and teacher policy variables, we can conclude that using the Fulanese proxies for SES will not significantly bias the estimates of the impacts of school variables on achievement.

(c) Rankings of school scores. Many accountability systems merely rank-order schools, as if uncontrolled mean scores were valid indicators of school quality. Even if one argues that such information is useful--for example, to citizens as an indicator of the level of attainment of a school's student body--it is not at all a good estimator of

the effect the schools themselves are having on students. And better estimates of that effect are useful for many purposes--to know which schools to study more carefully, to discover schools with severe problems, to provide parents with information that can be the first step toward local pressure for improvement.

How can we best rank schools as to their "value added"? If we had a valid theory of what variables constitute school success, we could compare those variables across schools. Unfortunately, we have no such theory. As a result, we can adopt a relative, residual measure of school's effects--one which credits schools with whatever variation is not explained by certain background and genetic factors that presumably operate prior to the effects of schooling.[1] The ranking that results is, on average, a "truer" ranking of relative school effectiveness given its student body's nonschool attributes.

There are, however, two problems. First, depending on which particular SES proxies are used, the amount of variation "left over" as potential school effects can change markedly.[2] Second, even with intricate SES measures, we will not be controlling for the true "home background;" insofar as we (necessarily) leave out important variables from our regressions, we bias our rankings of schools.

There are several tacks one can take to minimize the possible bias. First, after running various regressions, one can compare school rankings attained using one combination of SES measures and those using others. If the rankings shift wildly from grade to grade or across regressions, it would be best to ignore them or to report all of them; if

---

[1]Coleman et al. (1966) entered SES variables first in their regressions, thereby giving SES credit for any explained variance shared with school factors. This procedure explicitly assumed a causal priority. First home background has its effects, then schools do. This position has been redone to examine the extent of joint school-SES effects (Mayeske et al., 1969; Smith, 1972). For a further discussion of this technique of estimating school effects by residuals, see Barro (1970), Dyer (1972), and Klitgaard and Hall (1973).

[2]See, for example, the Michigan analyses in Klitgaard and Hall (1973) and Smith (1972, p. 257).

they are stable, one can have more confidence. Second, one can report rankings in a more "robust" form: for example, one could merely rank schools in terms of three categories, top third of the state given X, Y, and Z background characteristics, middle third, and bottom third. These points are discussed in the sections below on data analysis and data presentaion, respectively.

In conclusion, we should try to control for nonschool background factors as best as we can, confident that our estimates of Fulano's available school policy effects will not be significantly biased, but careful to check our rankings of schools for their sensitivity to the particular controls that are chosen. We are also confident that our "residualized" rankings of schools will be a better indicator on average of schools' relative "value-added" (in terms of achievement scores) than uncontrolled scores are. However, in both the analysis of the data and the presentation of results, we can take steps against people misusing such rankings.

In the case of Fulano, I recommended collecting the Urbanetics data for all units across which there is no transfer of students. For example, in rural areas where there is no bussing or open enrollment, use the building-level proxies; in areas where transfers of students are rampant, use the smallest units of analysis across which there is no movement (perhaps the district level, or some sub-district aggregate of buildings). Some schools would receive Urbanetics scores, then, that represented "their own" two-year-old proxies; others would get their district's proxies. This procedure would not estimate the "true" impact of SES, but it would eliminate some of the variability due to SES. As described below, the construction of an "SES composite" should proceed empirically, not entirely by deduction from perfectly specified SES variables; and in my opinion the Urbanetics proxies might contribute toward that composite, even if unevenly.[1]

---

[1]Much of the desirability of proceeding in this way depends on how the results are presented; see below.

The SES composite should also include the percentage of minority students and free lunches--if these proxies have predictive power. The inclusion of race would not be in tables (as, for example, the percentage of black pupils on each school), but as one among many SES predictors. Race would be used as a control rather than a descriptor. Such a procedure would avoid the potentially harmful political consequences of putting race "up front," while preserving that measure's well-known usefulness as a partial proxy for SES.

Additional SES data would be helpful, but I felt they would be too costly, politically and monetarily, for inclusion in the first year's report. I did urge that some variables (e.g., father's occupation and mother's education: a famous predictive pair) be obtained for the next year, either from questionnaires appended to each student's test or from a search through student files (a random sample of 30 from each grade would be enough, I think).

6. Data Analysis

I recommended the following methods of analysis, most of which can be done with "canned" programs. (Most, however, require a good programmer and a good statistician.)

For almost every elementary school, Fulano will have four reliable kinds of student scores:

3rd grade reading composite
3rd grade mathematics composite
5th grade reading composite
5th grade mathematics composite

If the six counties can be persuaded to give their tapes to the state, then at least the following statistics can be computed for each of the above four scores:

mean
standard deviation
% below statewide or nationwide 10th percentile
% above statewide or nationwide 90th percentile.

(If the district tapes cannot be obtained or processed, then perhaps only means and standard deviations could be examined statewide.)

One could regress each of the achievement score statistics for each of the grades against SES and IQ: a total of 16 regressions statewide. The first model is

$$y_{ijk} = f(\text{SES composite, distribution of IQ}),$$

where i is reading or math, j is grade 3 or 5, k stands for the achievement score statistic, and f is the functional form that yields the best fit to the data. These regressions would be based on a school-level tape with the following data: the 16 dependent variables; the SES proxies described in the previous section; and the distribution of IQ scores in the school (the following percentiles would be reported: 10, 25, 50, 75, 90).

The fitting process would require a good statistician, because it would basically be an inductive, exploratory process. One should select a random sample of, say, 400 schools. One should, for each of the 16 regressions, try to maximize the percentage of variation in the dependent variable that could be explained via some combination of SES and IQ (which could be viewed, if one liked, as another proxy for SES). A different combination would probably be necessary for each equation (although the regression equations for the four mean scores, for example, should look about the same) . The data fitting would be part art and part science; it might involve factor analysis and looking at residual plots; but most of it could be done on canned computer programs. The analysis would not be routine, but the programming basically would. I judged that both would be feasible tasks for the RSDE analytical staff, bolstered by an outside team of statisticians that the state was prepared to hire for the necessary period of analysis.

I suggested data analyses similar to those in Jencks et al.(1972) and other recent studies. Suppose one had estimated the best regression equation for all 16 Y_{ijk}. Each should be tried out on the full sample of data: if each equation explains about the same amount of variation as before and the regression coefficients stay about the same, the results look promising. One should then compute a school's residual (actual score minus the score predicted by the equation). It reflects the school's score after taking into account SES and IQ and is a proxy for the school's

own impact on the various achievement measures. To minimize the importance of random error for each type of score (mean, standard deviation, % $\leq$ 10%, % $\geq$ 90%), one could average the school's four residual values. Thus, suppose a school's mean residuals were:

3R μ = 2.5
3M μ = 0.6
5R μ = 1.1
5M μ = 1.6

Then its average mean residual would be the average of the four, or 0.1. (This case will probably be atypical of the amount of variability in the four scores.) Another example:

3R $\geq$ 90% = 3.2
3M $\geq$ 90% = 2.8
5R $\geq$ 90% = 2.4
5M $\geq$ 90% = 3.2

Average 90% = 2.9

Then one could list all the schools' average residuals for each of the four types of scores. Along each type of score schools would be divided into some small number of categories, say three to five, giving a scofe of 1 to those in the top third, a 2 to those in the middle third, and a 3 to those in the bottom third (and similarly for fifths).[1]

For example, the results might be:

	Avg. Residual Mean	Top, Middle, or Bottom Third?
School 1	2.6	1
School 2	-1.6	3
School 3	-0.1	2
School 4	+1.3	2

Similar tables would be produced for the other types of scores (standard deviation, % < 10%, % > 90%).

The point of using the four types of scores--as opposed to previous studies and systems, which have relied on the mean alone--would be to emphasize the multiplicity of goals in education.

[1] See also Dyer (1972).

The point of averaging the scores and then grouping them in this way would be to minimize random error and bias, as well as to indicate that all one can do is get a rough idea of effectiveness.

Districts can of course be analyzed merely by averaging their school's average residuals for each of the four types of scores.

Entering School and District Variables. One should also examine regressions with school and district variables included. The sign and significance of each school variable's regression coefficient would give a rough idea of its effect on the achievement measure, other things held constant.[1] Another 16 regressions could be performed, with similar averaging of each school measure's impact on the school's four scores of each different kind. Or the average residuals derived from the previous fit to SES and IQ alone could be regressed against the school and district variables. The choice would depend on resources available and the amount of multicollinearity between SES, IQ, and school proxies, but in any case the task would be fairly routine.

Tradeoffs among objectives. The move away from the mean to multiple dimensions of school outcomes offers another fruitful result: to help shift discussions by policymakers and citizens from one-dimensional levels of performance to equity, mobility, special programs, success with certain groups of students, and other educational goals. An important effect of an accountability system would be to educate policymakers and the public about the tradeoffs among goals. One way to emphasize the multiple and varied nature of educational objectives would be to examine the way the various achievement objectives correlate with one another. For example, does a higher mean imply a wider spread of scores? If a school does well with its students on the lower tail of the distribution, does it tend to do worse with the students on the upper tail? In short, what are the apparent tradeoffs between one desired outcome and another?

For example, one possibly useful analysis would be a correlation matrix relating the following achievement score statistics (objectives):

---

[1] See Cain and Watts (1970), Hanushek and Kain (1972), and Smith (1972), who prefers standardized regression coefficients.

uncontrolled mean score (achievement level), residual achievement score (achievement level relative to background), residual standard deviation (equalizing ability), residual percent of students below tenth percentile (success with underachievers), and residual percent of students above ninetieth percentile (success with overachievers).

Unusually effective schools. It was also recommended that the state search for unusually effective schools.[1] This suggestion was quite popular, as it had the prospect of showing success--and success would be useful both practically (to decisionmakers) and politically (to alleviate simplistic interpretations by the press that "nothing worked.") This suggestion was first made in November. By my January return, two district officials had produced regressions and scattergrams of residuals on their own, and were busy investigating the reasons for the success of several apparent outliers.

7. Presenting the Results

To serve two important objectives of evaluation--helping decisionmakers and minimizing adverse consequences--the presentation of the results of the accountability system is almost as important as the analysis itself. The guiding principles are the following:

(1) The report should have a summary and a brief and lucid main text. Most of the tables and all technical details should be relegated to appendices.

(2) The limitations of the exercise should be stressed, both in the text and in the way the quantitative results are presented.

(3) The emphasis should be placed on policy-relevant findings, rather than on mere descriptions.

The second principle has particular implications for the presentation of the regression and correlation analyses described above. Instead of reporting results in a continuous, cardinal fashion--with precise numerical estimates, statistics of significance, and so forth--these inexact results should be presented in categorical, almost qualitative form. Precision here is an illusion. The way we use the numbers ought to reflect this fact.

---

[1] See Klitgaard and Hall (1973).

TABLE 1

HYPOTHETICAL TABLE OF MULTIPLE MEASURES OF ACHIEVEMENT, BY DISTRICT

	MEASURES OF ACHIEVEMENT				
Districts	Achievement relative to background*	Achievement level	Equalizing capability*	% below 10th, relative to background*	% above 90th, relative to background*
A	3	1	[illegible]	2	3
B	1	4	3	4	2
C	5	4	4	4	2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Note: 1 = top 20% of districts in that category
2 = 60-80 percentile
3 = 40-60 percentile
4 = 20-40 percentile
5 = lowest 20% of districts in that category

* If rankings were very sensitive to the particular control variable employed, perhaps more than one such category of "achievement relative to background" could be reported, with an appropriate explanation.

TABLE 2

HYPOTHETICAL TABLE OF ESTIMATED EFFECT OF POLICY VARIABLES ON MEASURES OF ACHIEVEMENT

		Policy Variables	MEASURE OF ACHIEVEMENT 1. Achievement relative to background	3. Equalizing capability	5. % below 10th, relative to background	7. % above 90th relative to background
LEVEL	SCHOOL	Lower pupil-staff ratio	large positive *	large negative *	large positive *	positive close to zero
		Higher Teachers' education	negative	nagative close to zero	negative close to zero	positive close to zero
		Higher Teachers' pay	large positive *	positive close to zero	large positive	large positive *
		Higher Teachers' experience	negative close to zero	large positive	large positive *	large negative *
		Older age of building	positive close to zero	positive close to zero	negative close to zero	large negative *
		(etc.)		(etc.)		
	DISTRICT	Higher total cost per pupil	large positive *	large negative *	large positive *	negative close to zero
		Higher administrative cost per pupil	positive close to zero	positive close to zero	positive close to zero	negative close to zero
		Higher Instruction cost per pupil	large positive *	negative close to zero	positive close to zero	large positive *
		Higher personnel services cost per pupil	negative close to zero	large positive *	large positive *	positive close to zero

Key: Large positive = standardized regression coefficient greater than 0.1 (for example)
Positive close to zero = standardized regression coefficient between 0 and +0.1.
Negative close to zero = standardized regression coefficient between -0.1 and 0.
Large negative = standardized regression coefficient less than -0.1

Asterisks indicate findings that are statistically significant at the 0.10 level

TABLE 3

HYPOTHETICAL CORRELATION MATRIX AMONG STATISTICS OF SCHOOL ACHIEVEMENT

	1.	2.	3.	4.
1.	large positive	large negative	positive weak	large positive
2.		negative weak	large negative	large positive
3.			large positive	positive weak
4.				large negative

Key:

1 = achievement relative to background (residual mean)
2 = equalizing capability (standard deviation relative to background)
3 = success with underachievers (percent below national 10th percentile relative to background)
4 = success with overachievers (percent above national 90th percentile, relative to background)

Large positive = r = 0.50 or greater
Positive weak = r = 0 to 0.50
Negative weak = r = -0.50 to 0
Large negative = r = -0.50 or less

One example of a desirable way to present the results is the use of three-way or five-way groupings of schools along the various achievement statistics, as in Table 1.[1] Another example would be the presentation of the relationship between the policy variables and the achievement statistics as in Table 2. A third would be presenting the correlation analysis of the various achievement statistics as in Table 3.

In each case the precise results would appear in an appendix. The text, however, would discuss the results only in their qualitative significance. The analogy of the problem of significant digits in scientific experiments is appropriate: here our results are fraught with inexactness and conceptual limitations, and our use of numbers should be modified accordingly.

8. Conclusions

Policy evaluations, unlike much purely academic research, must concern themselves greatly with the benefits and costs of information for policy decisions, with the likely political consequences of the chosen methods of analysis and presentation, and often with the mere task of getting done.

---

[1] One should resist the temptation to concoct a grand measure, weighted sum of all the suggested statistics, and then to impose it on the evaluation process. Weighted sums assume mutual preferential independence, which probably does not hold (given most reasonable objective functions) for any of the measures. To take a comparable example: How one feels about income distribution probably depends on the general level of a country's income (Rescher, 1966, pp. 36 ff); on who the particular individuals are that fall beneath it. Similarly, assessing the intra-school spread is probably not advisable without considering the mean; and the existence of underachievers may bother one more if they are predominantly members of one ethnic or socioeconomic grouping. Although complicated alogrithms 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 decisionmaker (and each citizen) be his or her own judge. This tactic also avoids creating the one master list of schools, ranked from best to worst, that frightens local officials and is subject to endless misuse in popular discussions.

Consequently, what may be simply "technical" questions for the academic researcher can take on entirely different connotations for the policy analyst.

This case study suggested ways to improve the use of partial and inexact information in evaluating public education. Perhaps four general lessons can be drawn about making better evaluations, both in education and elsewhere:

(1) Employ multiple measures to reduce the misuse and opposition that accompany one-dimensional listings of schools and districts.

(2) Compare policy choices, not just schools and districts, which relate variables that state decisionmakers can affect to outcomes and reluctance of local officials to let their data be used.

(3) Search for unusually effective schools and districts, which are relevant to policy and are also likely to provide stories of success that can lessen the chance of misleading generalizations about the failure of education.

(4) Stress, not only in words but also in the way the numbers are presented, the uncertainty and inexactness of the evaluation's findings.

REFERENCES

Bailey, Stephen et al., "A Study of State Aid to Education in the Northeast," in Kirst, M.W. ed., The Politics of Education at the Local, State and Federal Levels, Berkeley, McCutchan, 1970

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

Bruno, James E., "Emerging Issues in Education: An Overview and Perspective," in Bruno, ed., Emerging Issues in Education: Policy Implications for the Schools, Lexington, Mass., Heath, 1972, pp. 3-28

Cain, Glen G., and Harold W. Watts, "Problems in Making Policy Inferences from the Coleman Report," American Sociological Review, Vol. 35, No. 2, April 1970, pp. 228-42

Coleman, James S., "Reply to Cain and Watts," American Sociological Review, Vol. 35, No. 2, April 1970, pp. 242-249

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

Cunningham, Luvern L., "Our Accountability Problems," in Sciara, Frank J., and Richard K. Jantz, eds., Accountability in American Education, Allyn and Bacon, Boston, 1972, pp. 78-91

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

_____, and E. Rosenthal, "An Overview of the Survey Findings," State Educational Assessment Programs, Princeton, Educational Testing Service, 1971

Hanushek, Eric A., and John F. Kain, "On the Value of Equality of Educational Opportunity as a Guide to Public Policy," in Frederick Mosteller and Daniel P. Moynihan, eds., On Equality of Educational Opportunity, Random House, New York, 1972. pp. 116-146

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

Kirst, Michael W., ed., The Politics of Education at the Local, State and Federal Levels, Berkeley, McCutchan, 1970

Klitgaard, Robert E., Achievement Scores and Educational Objectives, R-1217-NIE, The Rand Corporation, January 1974

_____, and G. Hall, A Statistical Search for Unusually Effective Schools, R-1210-CC/RC, The Rand Corporation, March 1973

Lessinger, Leon M., "Issues and Insights into Accountability in Education," in Bruno, James E., ed., Emerging Issues in Education: Policy Implications for the Schools, Lexington, Mass., Heath, 1972, pp. 229-249

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

Masters, Nicholas, et al., State Politics and the Public Schools: An Exploratory Analysis, New York, Knopf, 1964

Mosteller, Frederick, and Daniel P. Moynihan, eds., On Equality of Educational Opportunity, Random House , New York, 1972

Porter, Andrew C., and Garry L. McDaniels, "A Reassessment of the Problems in Estimating School Effects," paper presented at the American Association for the Advancement of Science, March 1974

Rao, Potluri, and Roger LeRoy Miller, Applied Econometrics, Belmont Calif., Wadsworth, 1971

Sciara, Frank J., and Richard K. Jantz, eds., Accountability in American Education, Boston, Allyn and Bacon, 1972, pp. 78-91

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

Wirt, Frederick M., and Michael W. Kirst, The Political Web of American Schools, Boston, Little, Brown & Co., 1972