Volume Title: Education as an Industry

Volume Author/Editor: Joseph T. Froomkin, Dean T. Jamison and Roy Radner, eds.

Volume Publisher: NBER

Volume ISBN: 0-88410-476-1

Volume URL: http://www.nber.org/books/jami76-1

Publication Date: 1976

Chapter Title: A Study of the Relationship of Instructional Process and Program Organization to the Success of Compensatory Education Projects in California

Chapter Author: Herbert J. Kiesling, Dean T. Jamison, Roy Radner

Chapter URL: http://www.nber.org/chapters/c4493

Chapter pages in book: (p. 249 - 290)

# 6

HERBERT J. KIESLING
Indiana University

# A Study of the Relationship of Instructional Process and Program Organization to the Success of Compensatory Education Projects in California

NOTE: The research reported in this study was jointly sponsored by the Carnegie Corporation of New York and the Rand Corporation. The manuscript as it now stands benefited from the careful readings and comments of Kent Anderson, Stephen Barro, Margaret (Polly) Carpenter, Stephen Carroll, Thomas Ribich, and Burton Weisbrod. The author also wishes to express appreciation to Harvey Averch, Ted Donaldson, and Marjorie Rapp of Rand for helpful advice, and profound gratitude to the Title I managers and other California school personnel who virtually without exception granted him their full cooperation. While the school district personnel must remain anonymous, persons who assisted from the California State Division of Compensatory Education are Tom Shellhammer, Alex Law, William McCormack, and Gerald Ryder, and from the Los Angeles County Department of Compensatory Education, William Turner and Norma Wilbur.

# I. INTRODUCTION

## Purpose of the Study

The study described in this report has two purposes. It is intended to advance our knowledge of compensatory education, especially with respect to issues of program design as it relates to the allocation of educational resources. In addition, I would hope that the study will prove valuable as a methodological experiment. Both objectives are important. The massive effort to overcome educational handicaps due to cultural deprivations authorized by Title I of the Elementary and Secondary Education Act of 1965 is one of the more important national social innovations of recent years. The program is costly, financed at an average of more than one billion dollars annually, and it is broad, aimed at all children coming from families officially classified as "poor."[1] Sponsors and proponents of the legislation have high hopes that it will be one significant way of drawing alienated poor and minority children into the mainstream of American life.

Despite its obvious importance, the program has been extremely difficult to evaluate. In large part, this has been because of the failure to develop a research methodology whose results would be useful to the policy maker requiring broad program evaluations. Thus, it would seem almost inevitable that any study making a contribution to our knowledge of the substantive questions involved will attempt new things in the methodology of educational research as well. Such an effort is made in this study; an evaluation of its success is left to the final section of the report, and beyond that, to the reader.

## Organization of the Report

Because of the methodological interests outlined above, the first section of the report includes a discussion of the place of this study in research relating to education policy. The following section deals with the steps taken to derive a model of compensatory education. It includes a description of past findings that suggest hypotheses to test, a description of the compensatory education process (which is used to generate testable hypotheses), and a discussion of the variables collected by questionnaire. A section is reserved for the findings, and the last section is devoted to a discussion of implications for further research contained in this report.

## II. METHODOLOGICAL CONSIDERATIONS

### General Background

In the past, there have been two fundamental approaches to policy-relevant evaluative research in education. To use descriptive terms developed by Averch et al. (2), they are the "process" and "input-output" approaches. The process approach, which characterizes most educational research of days gone by, is usually done in carefully designed experiments, often using experimental versus control-group methodology. These studies tend to have no standard method for reporting such student characteristics as socioeconomic background, attitudinal variables, and the like (beyond merely ascertaining that such characteristics are the same for both experimental and control groups). The criterion measure, or measure of performance, is whatever the researcher chooses, and there is very little consistency from study to study in terms of criterion measures, or if there is, they are usually measures which are of little direct interest to policy makers.[2]

In the input-output approach, quantifiable output measures, such as standardized objective test scores, are related to quantities of resource inputs, with some care being taken to account, at least roughly, for student differences in learning rate due to socioeconomic characteristics. This methodology overcomes the basic weaknesses of the process approach by using large samples with the same measure of output; but at the same time, it lacks the basic strength of process studies, which is the student-specific (or at least classroom-specific) nature of the analysis. The variables used have been aggregated by school buildings or school district (often for just one grade), and, further, they have not measured the personal traits of teachers or other school personnel but, instead, what Stephen Michelson has aptly termed their "objectified characteristics," years of experience, number of degrees, and the like.[3]

An important difference between the process and input-output methodological approaches is the statistical techniques they normally employ. Well-conducted process studies have traditionally compared the means of treatment and nontreatment groups for statistically significant differences. The emphasis has been upon finding that one treatment yields results that are "better than" another, without focusing greatly upon how much better the treatment group performed. Input-output studies have, on the other hand, used multiple regression techniques which, if assumptions underlying the statistical analysis are reasonably satisfied, have the important advantage of being able to trace functional

relationships between one variable and another, net of the effects of other variables entered into the regression equation. This advantage makes the approach potentially a more powerful statistical tool than the analyses of variance designs used in process research, although the latter are, perhaps, somewhat superior for studying interaction effects.

## III. BUILDING A MODEL OF THE COMPENSATORY EDUCATION PROCESS

The model used in this report is based upon a descriptive analysis of the compensatory education process and upon the findings of earlier studies of compensatory education programs.

### Description of the Compensatory Education Process

To begin constructing a model of compensatory education, it is useful to identify meaningful input variables through detailed analysis of the process sequence. In constructing the model for the empirical analysis, therefore, the starting point was a careful consideration of the problem of educating each child, including the organization, preparation, and actions which must be undertaken by the school from beginning to end in dealing with this problem.

In general, the "problem" of education usually begins with the realization that the pupil does not possess skills and attitudes which society wishes him to have. The education process is, of course, concerned with (effectively) dealing with the "problem" of lack of knowledge. A strategy for doing this must be picked, one which includes the training of instructional personnel, the planning of instruction, and the testing of results. In most traditional American education, preparation of instructional personnel occurs at the university, while planning and testing is the function of the individual teacher, who is not supervised to any great extent.

The education problem for children who are seriously underachieving should be viewed somewhat differently from that for normal children. Instead of "normal" lack of knowledge there is an "abnormal" lack of knowledge. The fact that the lack is "abnormal" implies that there exists some special reason for it, and the discovery of such reasons (diagnosis) becomes the important first step of compensatory education. Whether done explicitly or tacitly, formally or informally, the education of underachievers must begin with successful program diagnosis. Many states,

including California, realize this and require diagnosis as a part of Title I programs.

Successful diagnosis directly implies the need for proper prescription of instructional techniques which will effectively deal with the problems found in the diagnosis. The second step in the process is, therefore, prescription.

The third step in the process is to communicate the prescription for successfully overcoming the problem to instructional personnel, who, along with program managers and other decision makers, must execute the next step, which is to design and implement instructional techniques to accomplish the desired results. The final in-process step is to evaluate the success of the program. The evaluation step, especially if there is experimentation with different techniques, provides important feedback to all the other steps in the process.[4]

Although it is conceivable that a compensatory education program could get by without coordination of project members and effective leadership by the project director (e.g. in a project completely run by a reading specialist), in almost all instances observed by the author, teamwork of project personnel has been important. For example, even when the program is completely in the hands of a specialist, it appears desirable that he or she communicate periodically with the children's regular classroom teachers.

## Prior Findings

The research findings of two prior studies provide useful information about which aspects of the process just described should be contained in an input-output model. One is an earlier telephone interview study, conducted by me, of projects which were described by California State Compensatory Education personnel as highly successful.[5] Project directors were asked to describe their projects and to point out features which they considered central to program success. The second study (or, more properly, set of studies) was the painstaking review of project evaluations which was undertaken by Hawkridge and a number of associates at the American Institutes for Research (16, 17, 18). These authors first described the characteristics of studies which they could pinpoint as being successful. Then they found a number of projects which were quite similar to the successful ones in terms of objectives, basic program type, and pupil age, and attempted to ascertain which program components were associated with success and which with failure.

The findings for both of the studies just mentioned are briefly summarized in Table 1. They point to well-planned individualized instruc-

**TABLE 1 Factors Associated with Successful Compensatory Education Projects According to Studies by Hawkridge and Kiesling**

Hawkridge

*Preschool programs*

1. Careful planning, including statement of objectives
2. Teacher training in the methods of the program
3. Instruction and materials closely relevant to the objectives

*Elementary programs*

1. Academic objectives clearly stated
2. Active parental involvement, particularly as motivators
3. Individual attention for pupils' learning problems
4. High intensity of treatment

*Secondary programs*

1. Academic objectives clearly stated
2. Individualization of instruction

Kiesling

1. Individualization of instruction
2. Thorough planning and program coordination
3. Thorough in-service training of teaching personnel

SOURCES: David G. Hawkridge, G. Kasten Tallmadge, and Judith K. Lansen, *Foundations for Success in Educating Disadvantaged Children*; Final Report, U.S. National Advisory Council on the Education of Disadvantaged Children (Palo Alto, Calif.: American Institutes for Research in the Behavioral Sciences, 1968), pp. 19–20.

H. J. Kiesling, "California Compensatory Education Projects"; A Draft Report on the First Part of an Economic Analysis of Compensatory Education Projects in California. Working Note. (Santa Monica, Calif.: The Rand Corporation, 1970), p. 8.

tion as the key attribute of successful programs. Good in-service training is given prominent mention as well. Hawkridge and his associates concluded that motivation by pupils' parents was also important, at least at the elementary school level. These factors become, then, the program aspects which should be traced with special care in the analysis. In the next few pages the operation of compensatory education programs is considered in somewhat more detail as an aid to deriving workable variables.

## Individualized Instruction

### General Characteristics[6]

For purposes of this study, instructional techniques can be divided into two types: group and individualized. In group instruction, all members

of the class encounter the same set of experiences: they hear the same teacher lectures and comments by their peers, participate in the same exercises, and so forth. Students are required to learn at some minimum rate which is the same for everyone, although upward departures from the minimum are encouraged and rewarded.

When instruction is individualized, there is a relationship or interaction of the instructor directly with the individual pupils. Assignments are based on the individual needs of the student according to his ability, motivation, learning habits, previous attainments, and so forth. Sometimes pupils are given a degree of choice concerning curriculum in light of their own goals. Individualized instruction always involves individual diagnosis and testing to ascertain the pupil's problems and strengths. Sophisticated diagnosis may suggest the kind of instructional techniques which might best be used for each child or this may be ascertained in the course of instruction with experimentation. Pupil progress is evaluated continually.[7]

While individualized instruction is a complex process, this report will focus upon three key features that are central to its working. The first is the intensity of instruction, by which is meant the amount of instruction given to the pupil, the second includes the types of personnel and methods used to deliver the instruction to the pupil, and the third is the type of instructional materials used.

## Instructional Intensity

It is reasonable to expect that the amount of instruction given to pupils, other things being equal, would make a difference to program success. It is necessary to account for four sources of variations in treatment in measuring intensity: (1) the number of minutes per day per child, (2) the number of instructional sessions per week the child has, (3) the number of teaching personnel working with him, and (4) the number of pupils receiving instruction.

## Instructional Design

In American public schools, there is considerably more variation in the design of instruction for compensatory education than there is in that for normal education. Since design variations can be related to program quality, this makes the analysis for compensatory education considerably more interesting. Three kinds of personnel may be used: the regular classroom teacher who is released from part of her duties so she can give additional instruction to the compensatory education child; the trained specialist; and the paraprofessional, who is enlisted in support of either

classroom teachers or specialists. (Paraprofessionals are instructional personnel who are given on-the-job training and who do not have the required levels of formal education normally required for certification as a classroom teacher or as a specialist.) Also, the instruction itself is given either in the regular classroom or in some separate facility, usually a resource facility that has materials and supplies which will be discussed in the next section.

Since specialists receive training in individualized instruction techniques, it would be reasonable to expect that use of such personnel would yield better results. This view is supported by Guszak (12), who concludes that the disadvantaged child is best taught language skills by a diagnostic reading teacher who understands the variety of reading skills that exist and who can tailor instruction in skills to the individual while providing him with the emotional support that makes him wish to work and to achieve. Guszak also suggests that "the rank and file of teachers do not possess systematic knowledge of their reading skills program" (12, p. 363).

In light of the many criticisms of the role of certification in teaching effectiveness that have appeared in recent years,[8] it is also of great interest to analyze the role of the paraprofessional in the instructional process.

### Instructional Materials

Finally, it is likely that the type of instructional materials used will make a difference in the instructional effectiveness of individualized instruction. There is a long list of materials and equipment that are used in much greater depth for individualized instruction than in regular classroom instruction. Equipment commonly used includes recording sets with earphones, overhead projectors, films, film strips, controlled readers, and tachistoscopes. Nonmechanical teaching aids are used in even more profusion. These include word games of various kinds, flash cards, reading series, and encoding-decoding materials. In addition, most programs use considerable material made in class by the teacher or the students.

## Program Management and Coordination (or "Teamwork")

It is extremely difficult in a study with a small budget to get a good idea of the quality of program management. In this study, an attempt was made to examine program management indirectly by measuring program coordination or teamwork.

There are several benefits of teamwork. It makes possible the mutual reinforcement of goals through the dovetailing of instruction. It allows greater specialization. It encourages program personnel to share information about the problems and traits of individual children. Finally, it raises program morale. If the classroom teacher has no idea of what the specialist is doing, and no effort is being made to tell her, she may become somewhat suspicious and hostile, or, at least, indifferent. This attitude is quickly observed by the program children, and instructional effectiveness is harmed. If, on the other hand, it is obvious to the pupil that his teachers are working together, each with respect for the contribution of the other, he can respond to both without confusion.[9]

It is possible to use teamwork effectively in both group and individualized instruction, but the form that the teamwork takes in the two instances is somewhat different. In group instruction, specialization is limited mostly to areas of subject matter. Two instructors can engage in dialogue before the class, for example, or one instructor can cover material within his specialty one week, another the next, and so on. In individualized instruction, specialization and teamwork can be introduced into stages of the instruction process also. One person can diagnose the child's capabilities, another can give instruction, a third can supervise and counsel the primary instructor, and still another can evaluate the child's performance.[10]

The only program design in which it is possible to bypass most requirements for teamwork (and therefore management) is that which utilizes a highly trained and experienced specialist outside the regular classroom. He or she provides expert diagnosis, prescription, and instruction. If he or she has paraprofessional aid, it is possible for him or her to supervise them without help. And finally, he or she provides all of the ongoing evaluation and would only need a good clerk to tabulate the end of the year evaluation as well. Nonetheless, considerable teamwork is still useful in this kind of program. The specialist will often need additional diagnostic help from a psychologist or counselor. Outside evaluation is always helpful. It is almost always useful to inform both the principal and the child's regular teacher about the child's progress, needs, any special situations that require attention, and so forth. Thus, while it is possible to bypass a well-coordinated effort with this type of program, there might be a very real cost in terms of effectiveness in doing so.

Other program types require more teamwork. A program where the initial instruction is done by paraprofessionals in the regular classroom, for example, will require a specialist and/or a psychologist for diagnosis-prescription, a specialist to supervise aides, and much in-service training for aides and regular classroom teachers. A separate evaluator may be

required, as well as a full-time person as manager and coordinator—an individual whose talents are, of course, crucial to program success. If carefully designed, this type of program may be much less expensive than the "pure specialist" treatment described above.

There are organizational aspects to teamwork as well. Examination of formal and informal lines of authority in these programs would seem to be a most fruitful area for further research.[11] Questions to be explored would include whether the program manager has effective control over everyone in the program and whether he makes certain that the efforts of the various instructors with whom the program children come in contact are well coordinated.

Finally, there is room for teamwork in the evaluation phase of the program. With good individualized instruction day-to-day evaluation of the child's program is almost automatic. This may be done by the specialist working alone. But from the standpoint of broad policy objectives, good overall program evaluation may then be lacking.[12]

### In-Service Training

In my earlier telephone interview study, I was struck by the almost unanimous way in which respondents, upon being asked which aspect of their program did they deem most essential, answered "good in-service teacher training." In-service teacher training was mentioned in the conclusions of Hawkridge and his associates somewhat less often, although a careful rereading of a set of their key projects revealed that, indeed, the concept was present in virtually all of the successful programs and either specifically mentioned as absent, or not mentioned at all, in most of the unsuccessful programs.[13] These findings suggest that in-service training is quite important.

In-service training probably has a differential effect upon instructional personnel according to their background. For example, paraprofessionals may receive a considerable amount of in-service training but may nevertheless fail to provide instruction of the caliber of that provided by trained reading specialists (who presumably need much less in-service teacher training).

## IV. DATA COLLECTION

### The Sample

In the 1969–70 school year there were approximately 125,000 children in over 700 California Title I projects.[14] This study is based upon a sample

which represents about 6 per cent of these projects and 10 per cent of the pupils.

To insure comparability, only projects which employed the Stanford Reading Test were used. With this restriction, the sample was chosen on a stratified random basis, according to percentage of school pupils on AFDC (Aid to Families with Dependent Children), percentage black, and percentage with Spanish surnames. The sample is reasonably representative of the state in terms of pupil distribution although blacks are somewhat overrepresented and Anglos underrepresented in terms of projects.[15] The final sample includes 42 schools in 37 school districts all over California. There was a slight overrepresentation of schools in Los Angeles and Orange counties and underrepresentation of schools in extreme northern and eastern California for reasons of travel convenience. All but two of the interviews were given in person (otherwise on the telephone) and each interview took from 45 to 60 minutes.

There are two possible sources of bias in the sample. The first is due to the limitation of the Stanford Reading Test. While the Stanford was mandated by the State of California to be used in grades 2, 3, and 6 in 1969–70, only about 35 per cent of the Title I projects used it. It is widely thought to be a "difficult" test and perhaps districts which employ it have more than average self-confidence, which may be, in turn, based on actual high quality. On the other hand, the test was in fact mandated by the state, and districts which used it may be those which are efficient enough to use the same test for two chores or, perhaps, not ambitious enough to adopt what might be considered a more responsive test for the compensatory education program.

Another potential source of bias springs from the fact that only those projects that had readable reports were picked. (Every year about 15 per cent of all projects turn in reports which are not written well enough to allow meaningful interpretation.) If poor reports are the product of poor programs, there is obvious bias.

## The Questionnaire

The questionnaire was based directly upon the framework for studying the compensatory education process described above. Respondents were asked to report information on percentage minority and AFDC (these items could also be cross-checked from state sources), on instruction type, what aids were used, which personnel took part in instruction, size and length of classes, and class location. These data were double-checked since respondents were also asked to give schedules for the entire day of instruction personnel. Questions were designed to show who conducted diagnosis-prescription, to whom prescriptions were

communicated, which kinds of tests were used, and length of testing time. Similar questions were asked with respect to planning and in-service training. Finally, a series of questions were asked concerning lines of authority, including who decided, and who closely helped decide, on issues concerning hiring of program personnel, choosing program children, and a number of other program characteristics.

The questionnaire was pretested twice with analysis of problems and revision occurring after each pretest. It was designed to be given in person and to require only the responses of the operating manager of the school district Title I program if that person was well informed. In large school districts, however, it was necessary to interview both the building program manager and the district program manager. In numerous other instances, as well, information was obtained from others besides the primary respondent.[16]

The questionnaire is reproduced in its entirety in (26), pp. 37–47.

## V. VARIABLE CONSTRUCTION

### The Performance Measure

California compensatory education projects are required to submit performance data once yearly to the Division of Compensatory Education, including information concerning program objectives, instruments used, number of project participants by grade, project length, and frequency distributions of scores at the beginning and end of the treatment period. They are also asked to provide median pre- and posttest scores and the gain in grade equivalent by grade.

As mentioned above, some 35 per cent of all the projects which submitted reports to the state used the Stanford Reading Test. This made it possible to use the gains in standard grade scores on the Stanford test for the performance measure. Since the reports also include information concerning the specific objectives of these programs, it was possible to choose the sample only from schools which put as their major objective the raising of reading scores on standardized reading tests. This made it possible to overcome to some extent one of the comparability problems which has been noted in the literature, that of studying programs with different objectives (see McDill et al., [26]).

Two performance measures were used, ending score and gain in score per month of program duration (both in grade equivalents). The latter measure was used as an effort to consider separately from program

length the possibility that learning does not occur evenly over the length of the program, while the former measure was used because the use of gain scores has been criticized in the educational psychology literature. The measures were used for pupils pooled over grades 2, 3, 4, and 5, and for grade 3 alone, as that grade was the only one in which there were enough observations for meaningful analysis.

To keep this section from becoming overly long, the justification for these procedures, as well as the discussion of some other relatively minor problems concerning the performance measure, is reserved for an appendix.

### Beginning Level of Performance

It is conceivable that performance gain on standardized tests is not only a function of program treatment but also of where the children started. Often this relationship is positive: the pupils who start higher gain more.[17] If there is a test ceiling or "topping out" effect at work, however, the relationship might well be negative. In either case, proper specification of the model demands that the variable be included. As used in the estimating equations, the variable was coded as the number of months the children were below the national norm at the beginning of the program plus 20.0.

## Socioeconomic Variables

It is desirable to account for systematic differences in socioeconomic characteristics of pupil environments in order to assess the impact of the school program properly. Attempts were made to control for socioeconomic differences among pupils in two ways. First, respondents were asked to characterize the educational and occupational levels of the parents of their program children. This was, for several reasons, unsuccessful.[18] Second, a considerable amount of factual socioeconomic information was collected. Such data included the percentage of children in the school attendance area who were receiving aid for families with dependent children (AFDC) and the percentage of program children belonging to minorities.

### Pupil Mobility

Another characteristic that must be admitted to the analysis is the degree of mobility of program children. This may be a proxy for socioeconomic characteristics, since there are studies which show mobility to be positively related to low socioeconomic status (5). Mobility itself

can be injurious to program quality, of course.[19] Thus, even though a particular child stayed in the program all year, the quality of his instruction could be affected by the fact that his teachers are constantly bothered by the comings and goings of other children in the program.

## Instructional Intensity by Type of Instructor

As has been discussed above, the amount of instruction on an individual-equivalent basis was central to the analysis in this study. During the interviews, a record was made of how the pupils spent their project time, and this information was used to fashion the variables of individual-equivalent minutes spent with each child on a weekly basis by instructional personnel.

The variable as constructed allows for one measure to be constructed out of size of class, number of instructors, and length of session. As the variable was constructed, some allowance was also made for supervision time when the specialist, or classroom teacher, used one or more paraprofessional persons as assistants in actual instruction.

Here is an example of how the variable is constructed. If a single specialist sees groups of 10 pupils 30 minutes per day 5 days per week, the number of individual-equivalent minutes would be 15. (Thirty divided by 10 times 5.) If the specialist has one paraprofessional assistant for these 10 pupils, the number of individual-equivalent minutes for each pupil, ignoring supervision time, doubles. Since it is assumed that the specialist and the paraprofessional both lose 10 per cent of their time in the specialist's supervision of the paraprofessional, the number of individual-equivalent minutes for each pupil is not 30, but 27.[20]

There are of course three types of personnel used in instruction in the program, the trained reading specialist, the regular classroom teacher, and the paraprofessional. However, four types of instructor were used for constructing variables, with paraprofessionals divided into those assisting regular classroom teachers and those assisting reading specialists.

## Percentage of Instruction in the Regular Classroom

Considerable importance attaches to the relative effectiveness of supplementary instruction in the regular classroom as opposed to that given in a separate facility. If effective instruction could be given in the regular classroom, the cost would be much less and the regular classroom teacher could assume a more active part. She could also receive valuable in-service training in the course of her regular duties. On the other

hand, a specialist can give more undivided attention to children in a separate facility. We would expect to find a positive relationship between use of separate facilities and pupil performance, although this difference would probably be lessened in projects that have considerable teamwork and in-service training of regular teachers. The actual percentage of instruction given in the regular classroom was the variable used.

## Use of Educational Materials and Equipment

The possible importance of different types of educational materials and equipment was mentioned above. In the study, however, it was impossible in practice to determine the amounts of materials and equipment used. Thus, it was found that the essential characteristics of the lists of materials and equipment obtained for each program were virtually identical (at least to the untrained eye). To be sure, there were some differences in the amounts used, but these merely reflected the fact that there were more such materials in separate facilities and that reading specialists tended to use them more than regular classroom teachers. Because of this virtually complete overlap between percentage of instruction in the regular classroom and percentage of instruction given by the trained specialists, it was decided not to include a variable in the model for type of educational equipment used. It should be remembered, however, that any positive findings for percentage of instruction in the separate facility and instruction given by trained reading specialists must necessarily include in part a finding that there is possibly some return to the heavier use of such materials and equipment.

## Coordination and Leadership Variables

Several variables were used to represent program coordination. The simplest of these was hours spent in program planning per week. In the interviews, the respondents were informed what was meant by planning and by in-service training, and then asked how much of each took place. Since planning and in-service training are often difficult to separate, and also because there are problems with respondents' collective memories and with quantifying the length of informal discussions, both variables are probably subject to considerable measurement error.[21]

A variable to account for presumed weaknesses in lines of authority within the projects was also used. Teamwork should depend in part upon the degree to which all the principal actors in the project are subject to control by the same person. (Also, of course, it should depend

on whether he or she uses the control wisely.) The questionnaire was designed to discover not only the formal, but more importantly the informal, "chain of command." On the basis of the information collected, a dummy variable was constructed. It was set equal to unity when conflicts in direction and purpose were reasonably possible, and zero otherwise.[22]

One additional coordination variable was defined. Respondents were asked to identify the personnel who attended planning meetings and it was hypothesized that a well-coordinated program would routinely have more key personnel present at such meetings. The percentage of attendees who were considered key people became the variable.

### Use of Psychologists for Diagnosis

There was considerable variation in the amount of psychologist time used in the diagnosis and prescription phases of the programs. To test the hypothesis that intensive use of psychologist's diagnosis may be associated with better performance, a dummy variable was constructed on the basis of number of pupils per full-time-equivalent psychologist.[23]

## VI. FINDINGS

The model of school performance with the best explanatory power is presented in equation 1. All other variables discussed failed to add explanatory power to the model.

$$
\begin{aligned}
SCORE\ 25 = {} & 3.45 + 4.85\ PGMLENGTH^{*} + 0.86\ BEGIN\ 25 \\
& \quad (1.1) \quad (3.3) \qquad\qquad (7.4) \\
& - 0.013\ PCTMIN + 1.30\ SPECIEMS^{*} \\
& \quad (1.0) \qquad\quad (3.1) \\
& - 0.023\ PCTREGCR + 0.106\ TCHRPPIEMS \\
& \quad (1.7) \qquad\quad (2.3) \\
& + 2.07\ PLANHRS \\
& \quad (2.5)
\end{aligned}
\tag{1}
$$

$SEE = 1.84$; $F(7,34) = 21.32$; Corrected $R^2 = .78$

All of these models are weighted to correct for heteroscedastic error terms due to unequal numbers of pupils in each project.[24] The values given in parentheses are $t$ statistics and variables marked with an asterisk are transformed into their logarithms.[25] Variable descriptions are given in Table 2.

**TABLE 2 Means, Standard Deviations, and Description of Variables**

| Variable Name | Mean | Standard Deviation | Description |
|---|---|---|---|
| *SCORE 25* | 17.46 | 3.36 | Score at the end of program for students in grades 2, 3, 4, 5, in number of months relative to the grade level norm, coded such that the end score norm was 28.4 and the begin score norm was 20.0. |
| *SCORE 3* | 17.79 | 3.22 | Score at the end of program for students in grade 3, in number of months relative to the grade level norm, coded such that the end score norm was 27.8 and the begin score norm was 20.0. |
| *GAINSCORE 25* | 0.87 | 0.40 | Months gain on Stanford Reading Test per month of instruction, weighted average, students in grades 2, 3, 4, and 5. |
| *GAINSCORE 3* | 0.84 | 0.56 | Months gain on Stanford Reading Test per month of instruction, students in grade 3. |
| *PGMLENGTH* | 8.43 | 1.65 | Length of program in months, from pretest to posttest. |
| *BEGIN 25* | 10.88 | 3.25 | Months behind national norm of students at beginning of program, grades 2, 3, 4, and 5, plus 20.0. |
| *BEGIN 3* | 10.37 | 2.59 | Months behind national norm of students at beginning of program, grade 3, plus 20.0. |
| *PCTMIN* | 59.1 | 27.7 | Per cent of program children American Indian, black, and Spanish surname. |
| *SPECIEMS** | 18.0 | 13.7 | Number of individual equivalent minutes (IEMs)[a] per week taught by trained reading specialists. |
| *TCHRIEMS* | 16.3 | 10.1 | Number of IEMs[a] per week taught by regular classroom teachers. |
| *TCHRPPIEMS* | 8.8 | 8.4 | Number of IEMs[a] per week taught |

**TABLE 2 (concluded)**

| Variable Name | Mean | Standard Deviation | Description |
|---|---|---|---|
| | | | by paid paraprofessionals assisting regular classroom teachers. |
| *PCTREGCR* | 54.6 | 34.7 | Percentage of Title I instruction given in the regular classroom. |
| *PLANHRS* | 0.57 | 0.38 | Hours per week project personnel spent in planning meetings. |

[a] See page 262 for a description of individual-equivalent minutes.

Instruction both by specialists and by paraprofessionals assisting classroom teachers is related to pupil performance. For the paraprofessionals ten individual-equivalent minutes (IEMs) of instruction weekly are related to an additional month of reading performance. Specialist instruction shows a declining relationship with ten IEMs related to about 1.5 months of reading gain for the first ten or twenty minutes of instruction and then declining to less than one month of gain per ten IEMs beyond approximately 40 IEMs. The specialist variable was somewhat more statistically significant as well.

There is a small gain in performance when programs are conducted outside the regular classroom, although this variable is only barely significant at the 10 per cent level.

The only coordination-management variable which was related to performance was number of planning hours, with one hour per week of planning (which is more than most projects had) being associated with an additional 2.1 months gain. Causation cannot necessarily be inferred from the relationship, but it does suggest that some formal planning does indeed pay dividends. It is interesting to note that the in-service training variable, about which there were high hopes built on analysis of prior findings, always had the wrong sign and was never significant.

According to the variables both included and omitted from equation 1, no socioeconomic status (SES) variable is important. Of the variables not included, percentage of children with Spanish surnames had no explanatory power, while percentage black was weakly and insignificantly related to performance negatively. The percentage of children who moved, which can be considered as a proxy for one SES characteristic, was negative and usually yielded coefficients larger than their standard errors. The variable for percentage of children in the school attendance area on AFDC, which had been considered one of the more

meaningful SES variables, consistently displayed the wrong sign although it, also, was not statistically significant.

The percentage minority variable was somewhat collinear with amount of instruction conducted in the regular classroom ($R = .50$) and was somewhat more significant when that variable was not included in the model. To show this difference, equation 2 is a slightly different specification, with percentage of instruction inside the regular classroom being replaced by instruction by the regular classroom teacher.

$$\begin{aligned} SCORE\ 25 = {} & -4.89 + 4.47\ PGMLENGTH^* + 0.85\ BEGIN\ 25 \\ & (1.5) \quad (3.0) \qquad\qquad (7.0) \\ & - 0.023\ PCTMIN + 1.59\ SPECIEMS^* \\ & \quad (1.9) \qquad\quad (3.9) \\ & - 0.033\ TCHRIEMS + 0.090\ TCHRPPIEMS \\ & \quad (0.6) \qquad\quad (1.4) \\ & + 1.58\ PLANHRS \\ & \quad (1.9) \end{aligned} \tag{2}$$

$SEE = 1.91$; $F(7,34) = 19.53$; Corrected $R^2 = .76$

In this model the per cent minority variable is significant at almost the 5 per cent level. Specialist instruction becomes even more significant than before, but instruction by paraprofessionals helping classroom teachers loses some of its significance. Since more effective individualized instruction (including use of more specialized materials and equipment) is carried on in the separate facility, the first model, represented by equation 1, is undoubtedly much preferable to that in equation 2 on a priori grounds.

As the reader will recall, it was speculated that programs which depend almost exclusively upon reading specialists for their instruction might be expected to require less management and teamwork. To test this, the model was fitted to 25 projects which did not depend heavily upon specialist instruction.[26] The results are shown in equation 3.

$$\begin{aligned} SCORE\ 25 = {} & -7.65 + 5.33\ PGMLENGTH^* + 0.81\ BEGIN\ 25 \\ & (1.3) \quad (2.0) \qquad\qquad (6.1) \\ & - 0.011\ PCTMIN + 1.66\ SPECIEMS^* \\ & \quad (0.7) \qquad\quad (2.7) \\ & - 0.0063\ PCTREGCR + 0.109\ TCHRPPIEMS \\ & \quad (0.3) \qquad\quad (2.1) \\ & + 1.86\ PLANHRS \\ & \quad (1.4) \end{aligned} \tag{3}$$

$SEE = 1.89$; $F(7,17) = 13.89$; Corrected $R^2 = .79$

The importance of the planning hours variable is somewhat lessened instead of vice versa, and indeed this was true for all the other coordination and leadership variables as well. The hypothesis of better coordination in nonspecialist-dominated programs fails to be confirmed by the data.

Finally, because of the problems with respect to aggregating data from different grade levels mentioned above, the model was fitted to the 38 projects for which data were available for grade 3. The resultant equation, presented as equation 4, only manages to replicate the finding for the importance of specialist instruction, with the earlier significance of instruction of paraprofessionals helping classroom teachers and planning hours reduced to insignificance. This finding, therefore, introduced a note of caution into the interpretation of the meaningfulness of the latter two variables.

$$
\begin{aligned}
SCORE\ 3 = {} & \underset{(1.0)}{5.28} + \underset{(0.2)}{0.53}\ PGMLENGTH^{*} + \underset{(3.9)}{0.78}\ BEGIN\ 3 \\
& - \underset{(0.3)}{0.0060}\ PCTMIN + \underset{(2.6)}{1.60}\ SPECIEMS^{*} \\
& - \underset{(0.9)}{0.081}\ PCTREGCR + \underset{(0.7)}{0.048}\ TCHRPPIEMS \\
& + \underset{(0.6)}{0.76}\ PLANHRS
\end{aligned}
\qquad (4)
$$

$SEE = 2.59$; $F(7,30) = 4.08$; Corrected $R^2 = .37$

## Description of the Six Best Projects

The top-performing six projects in the study had pupil gains of at least 1.25 months per month of instruction. They averaged 1.5 months gain per month of instruction. It should be useful to outline briefly the characteristics of these six projects.

While four of the six projects had large amounts of instructional time for each pupil per week, the intensity of instruction in the other two was below average. It would appear, therefore, that large amounts of instruction are not absolutely necessary for good performance but are quite helpful.[27]

In five programs, a large proportion of the instruction was given by trained reading specialists. In the sixth, a paraprofessional who had three-years training by a specialist gave individualized instruction in a separate facility.

In the four projects in which the specialists employed paraprofessional aides, the amount of instruction given by the aide varied between

one-fourth and one-third the amount given by the specialist. In all projects the specialists gave instruction in small groups no larger than 10 students. Only two projects used classroom teachers and paraprofessionals in assistance of classroom teachers, and these two projects had large doses of specialist instruction besides. Four of the six programs had all instruction in a separate facility; the other two had half of their instruction in a separate facility.

There was no discernible trend among the six projects with respect to minorities represented. Three of the projects had a very high proportion of students belonging to minority groups; while in the other three, the percentage was quite small. Two projects had high percentages of black students and four had no blacks. Two projects had a high percentage of Spanish surname children. There was also considerable variation in pupil mobility in the six projects.

Concerning some other school variables, the number of pupils per full-time program manager in all six projects was quite low. On the other hand, the number of pupils per psychologist in the projects varied widely. The number of planning hours per week and the number of hours in-service training per week also varied quite widely. In all six projects almost all key people were present at all the planning meetings.[28] In several projects, the chain of authority appeared to be somewhat muddled, and therefore this variable does not seem to be very representative of high-quality programs.

In terms of geographical setting, the projects were all medium or small in size and were all either in rural or suburban settings. There were no large urban schools represented in the six top schools in the study.

To summarize the characteristics found in all of these highly successful projects, all six had small group instruction by specialists, high ratios of managers per pupil, and a consistently large percentage of key people present at planning meetings.

## Discussion of Findings

There has been wide commentary in the educational literature that compensatory education has failed; that there is no evidence to show that anything done in compensatory education programs is related to the performance of children from disadvantaged backgrounds.[29] The findings here with respect to the relationship of instruction by trained specialists to pupil performance, which maintain their significance no matter which of the meaningful subpopulations of these programs is chosen for fitting the model, clearly contradict this widely repeated set of findings. Instead, it supports the "reasonable hunch" of Guszak, based on work by

Turner and others, that the instructional procedures used by the diagnostic reading specialist are important. The evidence also suggests that instruction given by paraprofessionals helping regular classroom teachers may be effective.

Researchers who deal with disadvantaged populations often use 0.7 months per month of instruction as the "normal" rate of advance for these children, using traditional instructional methods.[30] The average gain in these projects was 0.87 months per month of gain. If the 0.7 figure is correct, the overall impact of the Title I money would be .17 months gain per month of instruction. For the projects which make heavy use of specialists giving individualized instruction, however, the gain is more. Increasing specialist instruction per child by twenty minutes per week should raise the average by at least .2 months, to a rate at which pupils would be slowly catching up. It would be dangerous to extrapolate the findings too closely in this way, but there is room for optimism based on the findings here.

Findings for the remaining aspects of the study are not nearly so positive, however. While it is true that the planning variable is significantly related to pupil performance in the main explanatory model used, the finding fails to hold up when the model is fitted to other meaningful subpopulations. Moreover, none of the other variables constructed to measure aspects of coordination and management were related to pupil performance at any time. With the possible exception of the finding for planning time, then, the general conclusion will have to stand that the strong hypotheses carried into the study with respect to the importance of coordination, teamwork, and management to program success, failed to be supported by the regression analysis. The descriptive results were somewhat more positive with respect to the importance of the amount of management input and to the percentage of key people who participated in planning sessions.

Whether the coordination variables failed because they represent reality, or because the variables are themselves too poor, remains to be seen in further studies. The latter possibility is considered highly likely although the very negative relationships found for some of the variables lead one to suspect strongly that the negative findings to some extent represent reality as well.[31] This suspicion is increased by the fact that nonspecialist-dominated programs had values for these variables which were even more negative in all cases than when the model was fitted to all projects. The same was also true for the in-service training variable, and the consistent null finding for that variable was something of a surprise and disappointment, considering all the rhetoric which I have heard in the past two years from program managers, directly and indirectly, concerning the importance of good in-service training. Perhaps

the problem was that we were not able to discriminate between good in-service training and poor in-service training, or perhaps the results are in part due to the fact that specialists (who are most effective in securing good results) do not require as much in-service training as other instructional personnel.

Proper discussion of the findings for program length and beginning score fall outside of my professional competence. Program length is related to performance, and the evidence suggests that more learning is done early in the program than later, since the variable fits the data much better when transposed into its logarithm. (This is also suggested by the negative coefficient for *PGMLENGTH* in equation 1-a in Appendix A.)

It is unfortunate that the model, when fitted to the grade 3 scores, did not replicate the findings for the teacher, paraprofessional, and planning variables which obtained in equation 1. In interpreting this difference, the question of how likely it is that the aggregation of data over different grade levels will lead to error immediately arises, and this question is discussed in more detail in Appendix A. I feel that the performance levels shown by the pooled grade data represent reality more faithfully than those which are for grade 3, but some readers may disagree after reading Appendix A.

If the pooled data findings are most representative of reality, the findings in the study are not all in one direction. Instruction by the classroom teacher with his or her paraprofessional (with instruction given by the paraprofessional counting most in this case) does in fact seem to be related to performance, to a degree about two-thirds as great as that for the trained specialist. If the significance level for the paraprofessional variable were the same, we could immediately draw some rather profound economic conclusions from this, of course, but since the confidence with which we can accept the paraprofessional finding is lower, it would be a dangerous extrapolation to make.

Finally, the difference in the relationship of socioeconomic status variables to performance in this study, as compared to other input-output type studies, should be noted. While most other studies have socioeconomic status (SES) as the quality most highly related to performance, no SES variable was significant here. Part of this can probably be explained by the fact that the other studies had pupil populations with wider variation in SES. This is even true with studies, such as those of Bowles (4) and Hanushek (14, 15) in which populations were restricted by race, since there were of course middle and high SES black or Spanish surnamed children present in their samples. This is the only input-output study which used low-status children exclusively. On the other hand, the variables used may have been inadequate. Even the

percentage of children in the school area on AFDC, upon which substantial hopes had been riding, completey failed to be related to performance. Much more sophisticated SES measures may be necessary for discriminating such things as verbalization in the home [see, for example, (5)], motivation, and the like. Yet, as indicated above, a procedure which depends on asking the child a straightforward question about these things is completely unacceptable for pedagogical reasons. It is perhaps surprising that the model explained as much of the variation in performance as it did, given the inadequacy of the SES variables.

## VII. CONCLUDING COMMENTS

This study is the first to attempt to assess compensatory education projects with input-output methodology. A single performance measure is used across all projects and an attempt is made to account for socioeconomic differences using multiregression techniques. As with other input-output studies,[32] the largest failure of this one is that the analysis is not student-specific, or even classroom-specific. However, an attempt was made to do some things which have not been done before in input-output studies, in that program organizational characteristics and instructional organizational strategies are related to pupil performance.

Since I lacked the necessary expertise to study the internal workings of the instruction, and also the necessary budget for employing highly refined techniques with organizational relationships, it is to be emphasized that the study is only a first step and that no more is claimed for it. It was hoped that this procedure might allow us a first, rather fuzzy look at the enigmatic inner workings of schools from the standpoint of input-output methodology, but only with respect to broad organizational patterns and not in a truly student-specific way. If this kind of methodology is to be pursued further, student-specific research will have to be added next.

It is certainly important for the cost effectiveness of the nation's educational research that wise heads carefully consider the payoffs to future research of the type undertaken here. It is by no means unanimously felt that such research will, in the future, yield results worth their cost. Thus, Alcaly, in commenting on the Hanushek study mentioned above, claimed that further studies of the same genre would probably not repay the cost (1). In commenting on an earlier version of the present paper, Ribich came to much the same conclusion (29). On the other hand, Weisbrod, in commenting on the same paper, said that there were probably increasing returns for many more studies of this kind (32).

If the approach does seem viable, the findings in the present study suggest several avenues for future work. The most pressing are expansion of the analysis of differences in instructional techniques and the inclusion of student-specific analysis. Individual students must be matched to individual teachers and treatments in large enough samples and with enough control for socioeconomic differences so that findings are statistically reliable. Secondly, much more careful thought will have to be given to program organization, coordination, and management. Some progress has been made in the past using role-analysis techniques in education, but further exploration must take place. Specialists familiar with organizational characteristics of large organizations, whether public or private, should be brought in to work on these questions. Finally, much more sophisticated work will have to be done to find meaningful socioeconomic variables.

## APPENDIX A

This appendix includes discussions of some statistical questions which were considered to be of insufficient general interest to be incorporated in the main text.

### Use of Gain Scores

As was indicated in the text, two performance measures were used in the empirical work done in this study. One of these was gain in grade equivalents per month of elapsed program time, and since there has been considerable criticism in the educational psychology literature on the use of gain scores because of the regression to the mean phenomenon (see Cronbach and Furby, [8]), only end-score was used in the findings presented in the text. Use of gain per unit of time elapsed does allow a direct look at the rate of learning over the length of programs, however, and moreover, a presentation of the model fitted to the gain variant should give some insight into the possible damage of using gain scores. The fitted equation which is similar to equation 1 in the text is therefore presented here as equation 1-a.

$$\begin{aligned} GAINSCORE\ 25 = {} & 0.85 - 0.031\ PGMLENGTH \\ & (3.5)\quad (1.3) \\ & - 0.015\ BEGIN\ 25 - 0.0016\ PCTMIN \\ & \quad (1.0) \qquad\quad (1.1) \\ & + 0.16\ SPECIEMS^{*} - 0.0032\ PCTREGCR \\ & \quad (3.3) \qquad\quad (2.0) \end{aligned} \tag{1-a}$$

$$+ 0.017\ TCHRPPIEMS + 0.25\ PLANHRS$$
$$(3.2) \qquad (2.6)$$

$SEE = .216$; $F(7,34) = 8.45$; Corrected $R^2 = .56$

Faster rates of learning appear to take place in the beginning of the program, although the program length variable is not statistically significant. It is also noteworthy that the overall findings one would infer from equation 1-a are very similar to those one would infer from equation 1.

## Pooling of Grade Data

Stanford reading scores were available for grades 2, 3, 4, 5, and 6 in various combinations from project to project. The number of valid observations for single grade levels varied from 38 in grade 3 to 15 for grade 5. Grade 3 was the only grade for which more than 50 per cent of the projects were represented. (A major reason for the large number of missing observations was that many projects changed test levels during the school year. This made their scores incomparable to the scores of projects which did not change levels.) Since achievement test scores are not necessarily comparable between grades (even when all scores are referenced to the norms by grade placement, as was done in this study) there is a possible objection to any procedure which pools data for different grades. On the other hand, if data were only used for the single usable grade, more than half of the performance data gathered in the study would have to be discarded. Discarding so much otherwise useful information is a step which should be avoided if at all possible.

The solution to this problem which was adopted was to use pooled data if no apparent differences could be found between grade results after analyzing grade differences statistically. The test used involved two steps. First, end-score was regressed against beginning-score for each grade to see if there were any discernible differences in this relationship by grade. There were not. Then, each grade was compared to grade 3, using a dummy variable for grade effect and covarying for beginning-score. (It was not necessary to covary for program length, since it was always virtually the same in the same school.) As an example of the procedure used, if there were twenty schools which had scores for both grades 3 and 4, the equation would have 40 cases and would be

$$SCORE = a_1 + a_2\ (BEGIN\ SCORE) + a_3$$

where $a_3$ is the coefficient of a dummy variable set equal to 1.0 if the observation were for grade 4 and zero otherwise.

The coefficients corresponding to $a_3$ for the four grade effects, with the $t$ statistics for their standard errors, are

| | Coefficient | $t$ |
|---|---|---|
| Grade 2 | − 0.08 | 0.42 |
| Grade 4 | − 0.09 | 0.33 |
| Grade 5 | 0.06 | 0.20 |
| Grade 6 | 0.42 | 1.52 |

Since the coefficient for the grade 6 effect was large and almost significant statistically, grade 6 scores for 440 pupils for 19 projects were excluded. All the other grades were retained and a weighted pooled average of both end-score and beginning-score was constructed.

What are the possibilities of this procedure leading to serious error? Differences in grade level effects could obtain because of different levels of resource inputs used at different grade levels, or because of differences resulting from test construction. Since we have statistical evidence that there is no difference between the four grades used, the kind of errors that could remain in the presence of this null finding would be offsetting errors, that is, increased resources might be used at a grade in which this factor is offset by the effect of test construction which biases gains downward. However, considerable care was taken in the interviews to check for differences in inputs by grade level, and there were not many instances in which they obviously differed (this is especially true with respect to grade 2, somewhat less true, perhaps, with respect to the findings for grades 4 and 5).

I doubt that this pooling procedure has led to serious error. Readers who disagree will have to use the findings presented in equation 4 and disregard the rest.

## Other Minor Problems in Constructing the Performance Measure

There were a number of relatively minor problems to overcome in using the Stanford Test Scores in this data set. First, it was found necessary to use the median performance scores as the measure of central tendency, since some projects failed to include frequency distributions in their reports. (Such frequency distributions would have been required to compute means.) This procedure allows for some bias, but careful investigation showed that the difference between mean and median grade equivalents (many districts reported both) were nonexistent or negligible.

A second problem arose because it was not possible to obtain summary scores for individual schools from some of the school districts. Twenty-two of the 42 school projects fell into this category. Half of the 22 had district reports in which the school project being studied accounted for less than half of the pupils covered in the report. The method used to attempt to overcome this potentially serious data problem was to request the respondent to choose a school that was "closest to the district average" in performance. Some such choice was usually possible, and since district-evaluation personnel often have a good feel for the performance levels of their project schools, the error introduced by the mismatch was probably lessened considerably by this procedure.

It may be of interest to some readers to see the model fitted to only those 31 projects where the mismatch problem was—in terms of percentages, anyway—relatively minor. This is done in equation 2-a.

$$
\begin{aligned}
SCORE\ 25 = {} & -3.32 + 4.35\ PGMLENGTH* \\
& \quad (0.7) \quad (1.9) \\
& - 0.206\ BEGIN\ 25 - 0.0040\ PCTMIN \\
& \quad (1.7) \qquad\qquad (0.2) \\
& + 1.48\ SPECIEMS* - 0.022\ PCTREGCR \\
& \quad (3.1) \qquad\qquad (1.3) \\
& + 0.089\ TCHRPPIEMS + 0.80\ PLANHRS \\
& \quad (1.8) \qquad\qquad (0.8)
\end{aligned}
\tag{2-a}
$$

$$SEE = 1.77;\ F(7,23) = 4.22;\ \text{Corrected } R^2 = .43$$

Except for the less significant *PLANHRS* variable the equation is not greatly different from equation 1.

Finally, there was a problem with respect to the question of competing program outputs. The California Division of Compensatory Education requires that Title I projects teach both mathematics and reading. It was not possible to obtain comparable achievement data on mathematics for 18 of the 42 projects,[33] however, and with this many missing observations it was simply not feasible to study mathematics programs directly. Instead a careful attempt was made to limit the study to resources going into reading.

## Weighting

A problem well known to econometricians concerns the fact that regression equations fit to sample populations, where the expected error terms from properly specified models are not the same size along some important dimension of the analysis, are not efficient. That is to say, other estimators can be found for which there is less error variance. There is

one dimension in educational analysis like this study where such expected error variance must surely differ, and that is program size. This is because mean scores of groups of pupils are used and the expected error variance of means of small groups is greater than those for large groups, as everyone who has studied sampling theory knows.

In studying educational projects of this kind, there is one additional quirk to the analysis which has not been pointed out before in the educational input-output literature.[34] There are two potential sources of randomness, a program effect applying to each student in the program, and a random effect which differs for each student and which arises from the vagaries of achievement testing. In symbols $u_{ij} = v_i + e_{ij}$ where $u_{ij}$ is the stochastic term for the $j$th student in the $i$th program, $v_i$ is the effect of the $i$th program, and $e_{ij}$ is a random term. The variance of the average test score across all students in the $i$th program depends on the number of students (size of program), because the sum of $e_{ij}$ depends upon the number of students. The variance $v_i$ due to program effects may or may not depend on size of program. (In point of fact, I would suspect that it does, since the law of large numbers works with teacher's effects and the like, as well as with pupil performance on tests.) If $v_i$ is independent of size of program, the question then becomes: "How much of the total error term $u_{ij}$ varies by program size and how much does not?" If a large percentage did not vary, it might be more correct not to weight, or to use only a partial weight.

It should be possible to gain some insights concerning the propriety of weighting fully merely by performing the well-known test for heteroscedasticity. The projects were divided into four groups of 10, 11, 11, and 10, respectively, ranged by sample size, and the variance of the error term multiplied by a constant was computed for equation 1-a. The result was as follows, where $N$ equals the number of pupils in the project whose scores were averaged.

| $1/N \times 1000$ | Variance $\times$ 100 |
|---|---|
| 5.8 | 36.4 |
| 13.1 | 38.4 |
| 23.3 | 49.4 |
| 54.2 | 129.6 |

Variance obviously increases consistently with decreased sample size. If a regression line of variance is hand fitted to $1/N$, the resulting line has a steep slope and an intercept fairly close to zero. This seems to indicate strongly that full weighting on the basis of sample size is proper.

## NOTES

1. The Report on Title I for the 1968 fiscal year gives the number of children in poor families as 7,700,000 (29, p. 66). Of these, 89 per cent are in schools which receive Title I aid, and about 52 per cent are participating in some form of Title I program (29, p. 14 and p. 87).
2. For example, many of the criterion measures of teacher performance are ratings by their superiors as to the quality of their performance. There is seldom any effort to obtain correlations of ratings by superiors and actual classroom performance.
3. Two exceptions to these remarks must be noted. One Rand-sponsored study, by Hanushek (15), has matched pupils in grades 2 and 3 with their teachers. Also a number of studies, including those based on the Coleman Report and the Hanushek study just mentioned, have had variables for teacher performance on a simple verbal abilities test.
4. See Rapp (28).
5. The success criterion used was gains in cognitive reading tests which approached two times what was considered "average" for low socioeconomic status (SES) children. See Kiesling (23).
6. The following discussion has benefited greatly from the series of monographs on the subject of individualized instruction which have been written at the Far Western Regional Laboratory (9).
7. Despite what may seem logical in the matter, class size for individualized instruction is *not* necessarily smaller than that for group instruction. It is the teaching technique, not the class size that is important. Group instruction, with virtually no individualized instruction at all, could be carried on (and often is, for example, in graduate courses) with classes of four or five. Individual instruction techniques often include giving the child a short assignment and sending him off to do it. A good specialist instructor can probably give individualized instruction to 20 children at once. In actual practice, it is probably seldom that either type of instruction is given in pure form. For example, if a specialist instructor worked directly with an instructional aide, it was assumed that 10 per cent of the time of both was spent in discussion between themselves for reasons of supervision, and not in actual instruction of children.
8. See Kiesling (25, p. 34).
9. The individualized instruction that a pupil receives as part of the program is likely to be a pleasant experience, because he feels that someone cares enough to get to know him personally and to be his friend. If he feels that his regular classroom teacher is highly sympathetic to his compensatory instruction, he may relate his pleasant experience to his regular school program, resulting in a much improved attitude to all of his schoolwork.
10. Some of the instruction can be performed separately in group instruction, too. Separate people can supervise and evaluate, for example. In practice, however, this is seldom done.
11. Some work along these lines has been done. See, for example, Halpin (13), or Katz and Kahn (22).
12. For a good discussion of how this can be done, see Rapp (28).
13. Hawkridge, *et al.*, (18).
14. It should be noted that two schools in the same school district are considered to be two projects.
15. This is because a disproportionate number of blacks were in a few large schools.
16. Often as I conducted my interview and came to a section of questions which the respondent did not feel competent to answer, he or she would get me a quick

appointment with someone who knew the answers (or at least give me his name and telephone number for a telephone query later) or pick up the telephone and call someone to find out while I waited. An advantage of giving the questionnaire in person is that it is quickly ascertained to the mutual agreement of both interviewer and interviewee when the latter is weak with respect to knowledge of some program aspects. As noted in my acknowledgment, I received a degree of cooperation from almost all school personnel which I think rather amazed my colleagues at the Rand Corporation.

17. In an earlier study of mine, gains in performance from grade 4 to grade 6 were highly correlated with score in grade 4. See (24).
18. Data concerning family characteristics which might bear upon pupil motivation are simply not collected. The reason for this is understandable. Many children in Title I programs come from homes which, unfortunately, have characteristics about which they feel embarrassed. Many program instructors feel that merely asking children questions concerning their home environment causes an adverse effect upon pupil morale and pupil achievements.

    It seems to me that it should be possible, however, to overcome this problem by administering instruments or questions to the children which might, directly or indirectly, assess such characteristics as amount of verbalization in the home, and so forth, without directly embarrassing the child if there is some problem. The use of one such test is described in (6).
19. It should be noted that mobility does not directly affect the performance outcomes, since test scores were reported by the projects only for pupils present both at the beginning and the end of the program.

    The question that was asked to obtain mobility rate was: "What percentage of those children who were initially placed in the program at the beginning of the program year were still in the program at the end of the program year?"
20. The convention used was to deduct 10 per cent of the instructional time of supervising teacher and paraprofessional for each of the first two paraprofessional aides, and 5 per cent for each aide after that.
21. As was explained to the respondents, planning was defined to include the kinds of topics and skills program personnel should be covering during the coming week or weeks for individual children (by name). By in-service training was meant explanations concerning why project personnel should take various educational steps, how and when a certain skill requires that another kind of skill be taught immediately prior, and so forth. Demonstrations concerning classroom techniques suited to teaching skills which the program leaders desire to be taught are also included.
22. An example of the "no conflict" situation would be where the program is directed by an Assistant Superintendent with line authority who is not too busy to devote a reasonable amount of time to the program. Thus, no coordination problem need ever arise: all personnel concerned, including specialists, building principal, and so forth, are directly responsible to the Assistant Superintendent.

    A majority of the actual programs were included in the "conflict possible" category, however. Often, for example, the program director has a rank equal to the building principal and has no "line" authority. The Director might supervise the specialist within a given school, while the building principal supervises the classroom teacher and paraprofessionals. The success of such a program depends crucially upon how closely the Director and the building principal cooperate. Even if these two individuals are good friends, chances are that the effects of the specialist and regular classroom teacher may not be well coordinated. At least, this is our supposition. A variation of this pattern exists when a person has the control but has too many other duties to effectively use it to coordinate the program.

23. There were very few projects which had a ratio of pupils to full-time-equivalent psychologists near 1000:1. Since most projects fell either clearly above or below this figure, if the ratio was below 1000:1 the dummy variable was set equal to unity and if above, to zero.
24. Weighting is further discussed in Appendix A.
25. For 34 degrees of freedom, significance levels are: 5 per cent 2.0; 1 per cent 2.7; .1 per cent 3.5.
26. The criteria used in making the distinction were that more than half of total instruction was accomplished by specialists together with paraprofessionals assisting specialists; and at the same time, more than half of all instruction took place in a separate facility.
27. The average number of IEMs for all 42 projects was 44 and the two projects mentioned as below average had 37 and 25 IEMs respectively. The difference in instructional intensity between the best and worst projects is striking, however. The average number of IEMs for the six best projects, including the two just mentioned, was 70. The average for the 10 worst projects, which had an average gain of about .4 months per month of instruction, was only 32. The difference in the amount of instuction given by trained specialists is even more striking: 30 IEMs in the best projects as opposed to 12 in the worst.
28. This was not true in the ten worst projects, where the per cent of key people average was 75. It is notable that, in these ten projects, for those in which the percentage of key people present was large, the actual planning time was small.
29. To cite only two: "Compensatory education has been tried and it apparently has failed"—Jensen (20, p. 2). "Negative residual gain-scores for most 'participating' groups in all grades seem to indicate that even when a lower 'starting point' is considered, participants did not progress at the same rate as nonparticipants"—Glass et al. (10, Chapter 6, p. 148).
30. The figure found in the Coleman Report was that disadvantaged children who reach grade 12 are about 3 grade levels behind. This would imply a figure of .75 months per month of instruction for those *who do not drop out.*
31. A cynical explanation, which I would be inclined to reject, is that all projects had uniformly bad management so there was nothing good to measure. I would also be inclined to reject the opposite explanation that all projects had management that was uniformly good.
32. Except Hanushek's (15), which was classroom specific.
33. Some districts did not include mathematics in their annual reports and others did not use the Stanford mathematics tests.
34. I owe this point to Joseph Newhouse.

## REFERENCES


1. Alcaly, Roger. "Discussion." *American Economic Review: Papers and Proceedings* 61 (1971): 301–302.
2. Averch, Harvey, et al. *What Do We Know About Educational Effectiveness?*, R-956-PCSF/RC. Santa Monica, Calif.: The Rand Corporation, 1972.
3. *Annual Report.* Officers of Local Assistance, Department of General Services, State of California, Sacramento, 1970.
4. Bowles, Samuel. "Educational Production Function." In W. L. Hansen, ed., *Education, Income, and Human Capital.* New York: NBER, 1970, pp. 11–61.

5. Bogue, Donald J. *Principles of Demography.* New York: John Wiley and Sons, 1969, pp. 770–771.
6. Carver, Ronald P. "Use of a Recently Developed Listening Comprehension Test to Investigate the Effect of Disadvanterent on Verbal Proficiency." *American Educational Research Journal* (Mar. 1969): 263–270.
7. Coleman, James S., and Karweit, Nancy L. *Measures of School Performance,* R-488-RC. Santa Monica, Calif., The Rand Corporation, 1970.
8. Cronbach, Lee J., and Furby, Lita. "How We Should Measure 'Change'—Or Should We?" *Psychological Bulletin* 74 (1970): 68–80.
9. Far West Laboratory for Educational Research and Development. *Individualized Instruction Information Unit.* Berkeley, Calif., 1971.
10. Glass, Gene V., et al. *Education of the Disadvantaged: An Evaluative Report on Title I, Elementary and Secondary Education Act of 1965, Fiscal Year 1969.* Boulder: University of Colorado, 1970.
11. Gordon, E. W., and Wilkerson, D. *Compensatory Education for the Disadvantaged.* New York: CEBCO Standard Publishing, 1966.
12. Guszak, Frank J. "The Diagnostic Reading Teacher for the Disadvantaged Child." In J. L. Frost and G. R. Hawkes, eds., *The Disadvantaged Child.* Boston: Houghton Mifflin, 1970, pp. 361–378.
13. Halpin, Andrew W., and Croft, Dan B. "The Organizational Climate of Schools." St. Louis, Mo.: Washington University, 1962, processed.
14. Hanushek, Eric A. *The Education of Negroes and Whites.* Ph.D. diss., M.I.T., 1968.
15. ———. *The Value of Teachers in Teaching.* RM-6362-CC/RC. Santa Monica, Calif.: The Rand Corporation, 1970.
16. Hawkridge, David; Chalupsky, Albert B.; and Roberts, A. Oscar H. *A Study of Selected Exemplary Programs for the Education of Disadvantaged Children.* Final Report, U.S. Department of Health, Education, and Welfare. Palo Alto, Calif.: American Institutes for Research in the Behavioral Sciences, 1968.
17. Hawkridge, David; Campeau, Peggie L.; DeWitt, Kathryn M.; and Trickett, Penelope K. *A Study of Further Selected Exemplary Programs for the Education of Disadvantaged Children.* U.S. Department of Health, Education, and Welfare. Palo Alto, Calif.: American Institutes for Research in the Behavioral Sciences, 1969.
18. Hawkridge, David; Tallmadge, G. Kasten; and Larsen, Judith K. *Foundations for Success in Educating Disadvantaged Children.* Final Report, U.S. National Advisory Council on The Education of Disadvantaged Children. Palo Alto, Calif.: American Institutes for Research in the Behavioral Sciences, 1968.
19. Jaeger, R. M. "The 1968 Survey on Compensatory Education." Paper presented at the American Educational Research Association Meeting, 1969, Los Angeles, Calif.
20. Jensen, Arthur R. "How Much Can We Boost I.Q. and Scholastic Achievement?" *Harvard Education Review* 39 (1969): 1–123.
21. Johnston, J. *Econometric Methods.* New York: McGraw-Hill, 1960.
22. Katz, Daniel, and Kahn, Robert L. *The Social Psychology of Organizations.* New York: John Wiley and Sons, 1966, Chapter 7.
23. Kiesling, Herbert J. "California Compensatory Education Projects: A Draft Report on the First Part of an Economic Analysis of Compensatory Education Projects in California." Working Paper. Santa Monica, Calif.: The Rand Corporation, 1970.
24. ———. *The Relationship of School Inputs to Public School Performance in New York State.* P-4211. Santa Monica, Calif.: The Rand Corporation, 1969.

25. ———. "Multivariate Analysis of Schools and Educational Policy." P-4595. Santa Monica, Calif.: The Rand Corporation, 1971.
26. ———. *Input and Output in California Compensatory Education Projects*, R-781-cc/rc, Santa Monica, Calif.: The Rand Corporation, 1971.
27. McDill, Edward; McDill, Mary; and Sprehe, Timothy. *Strategies for Success in Compensatory Education: An Appraisal of Evaluation Research*. Baltimore: Johns Hopkins University Press, 1969.
28. Posner, J. "Evaluation of 'Successful' Projects in Compensatory Education." U.S. Office of Education, Office of Planning and Evaluation, Occasional Paper No. 8, 1968.
29. Rapp, Marjorie. *Evaluation as Feedback in the Program Development Cycle*. P-4066. Santa Monica, Calif.: The Rand Corporation, 1969.
30. Ribich, Thomas I. "Comment on a Paper by H. J. Kiesling." Conference on Education as an Industry, National Bureau of Economic Research, Chicago, June 4–5, 1971. [This volume.]
31. *Statistical Report for Fiscal Year 1968: A Report on the Third Year of Title I*. U.S. Office of Education, Department of Health, Education, and Welfare, Washington, D.C.: GPO, 1969.
32. U.S. Office of Education, Department of Health, Education, and Welfare. *Evaluative Report on Title I, Elementary and Secondary Education Act of 1965, Fiscal Year 1968*. Washington, D.C.: GPO, 1970.
33. Weisbrod, Burton A. "Comment on a Paper by H. J. Kiesling." Conference on Education as an Industry, National Bureau of Educational Research, Chicago, June 4–5, 1971. [This volume.]

# 6 COMMENTS

## Thomas I. Ribich

University of North Carolina, Chapel Hill

Before getting into the specifics of the paper presented by Kiesling, I should like to state my general position on the matter of educational production-function studies. Briefly, I'm a grouch. Part of the reason for my present predisposition can be found in the sort of technical problems and social-strategy arguments posed by Henry Levin in the opening paper of this conference, and part is present in the line of criticism developed by John Brandle more than two years ago at another conference on the economics of education sponsored by the NBER. But there are other things as well. Some

**EDITORS' NOTE:** A number of the objections raised to Kiesling's original paper in Ribich's discussion have been taken into account in the version of Kiesling's paper which appears in this volume.

of them are very general, some seemingly incidental and just barely germane. I would like to mention them briefly, if, for no other reason, than to make it clear that the specific criticisms I have of the paper under consideration may be due mainly to a bias against the genre rather than hostility toward this particular study.

To start with what may be the most incidental of all, I am a teacher as well as someone who does research about the economics of education. As a teacher, associating regularly with other teachers of the same general subject matter, I am convinced there are just about as many valid ways to teach a course as there are teachers. There also seem to be about as many ways to run an economics department as there are department chairmen. Production function studies have not, and probably never will, change my mind about that. Now, perhaps it is a different story when it comes to education and the administration of education at lower levels, but I tend to doubt it. There are different teaching styles and different administrative styles, based largely on personality and the like, and they tend to employ various "resource inputs" with greatly different degrees of efficiency. I believe that this view is widely shared, though rarely (if ever) articulated in "professional" discussions of educational production function. Nevertheless, such studies (especially those like Kiesling's, which go into considerable detail on instructional technique) will, I think, continue to be unpersuasive to a good many of those on the frontlines because of this basic problem.

Second, I cannot help feeling that a lot of the educational production-function studies are forced flowers, stimulated in a hurry as a result of the startling findings of the Coleman Report, and of several other studies appearing about the same time, indicating that educational inputs as a whole were dwarfed by socioeconomic variables when it came to explaining educational outcomes. The consequent search for educational inputs that do have a statistically significant effect on educational outcomes has turned up some apparently interesting results, but the manipulations and strainings of statistical tools required to make sense out of very imperfect data has led to complexity so bewildering as to leave the uninvolved onlooker (like myself) deeply uncertain about whether there is any meaning at all in the tables repeatedly confronted. On top of that, much of the analysis and controversy has been conducted in a vaguely tense atmosphere having to do with preserving or impugning the honor of the educational establishment. That, plus the sheer tedium of wading through mounds of theoretically ungirded statistical analysis has given the whole subject (to me at least) the emotional content of prolonged trench warfare.

There are a few other things I could mention—my feeling that the theory economists are presently getting into, in order to shore up the statistics, is of an "engineering" sort, for which economists have no special insight and expertise, and my hunch is that important truths and the means of their discovery should be simpler and prettier than the kinds of things we are turning up and the ways we are employing to do it. But I think what I've said already should be more than enough to certify me as an authentic grouch, and the specific criticisms I turn to now should be considered accordingly.

To anticipate by paraphrasing the title of Jim Bouton's new volume, "I'm glad the author is not going to take it personally."

The following has only a rough logical sequence, and the points are clearly separable, so I'll rely on the crutch of enumeration. The criticisms are not mortal blows but rather a pointing out of ambiguities and ironies, not unlike the kind that can be enumerated for most studies of this sort. There are at least a few special twists, however.

1. The statistical proceedings in Kiesling's paper make me no less uneasy than do most other papers on educational production functions. He states that he has gone to a small survey format in order to duck some of the problems he views as inherent in a "massive survey" approach or present in the context of a small "controlled" experiment. Yet in the course of explaining his methodology and his results, we are confronted with one instance after another of arbitrary assumptions, apologies for the crudeness of the data, and difficult-to-unravel statistical conundrums. One almost gets the feeling that the small survey approach manages to combine the problems that exist in the small experiments and large surveys more than it succeeds in slipping between the horns of the dilemma.

2. It is hardly comforting that the input which by calculation appears to be twice as cost effective as anything else—i.e., "better teamwork"—is not even statistically significant. Kiesling states at one point that his "intuitive judgement" is that "the magnitude of returns to expenditures on management (and thereby teamwork) is not overstated." But in summarizing his earlier regressions, upon which the cost-effectiveness estimates depend, he notes that "the effect of better teamwork between program personnel seems to be positive, although it is not possible to ascertain the magnitude of the effect." Moreover, Kiesling avoids using the teamwork variables in the construction of his isoquants later on. The issue is never resolved, and the basis for resolution is never spelled out, though it seems apparent that it is buried in the multiple ambiguities of the imperfect measures of teamwork, the question of how much teamwork can be deliberately encouraged, and the arbitrariness of the cost estimate.

3. Taking a brief look at the isoquant analysis, the diagram summarizing the results suggests strongly to me that most schools are using too many specialists. Surely the most technically efficient firms are by and large in that position if the efficiency frontier is to be taken seriously. Yet the regression analysis, converted into cost-effective terms, suggests that *not enough* specialists are being used, in that they could be substituted for individual instruction time provided by regular teachers with the result that test-score gains will rise without an increase in costs. Kiesling seems to give precedence to the isoquant results noting that "while average relationships are the proper ones to explore for making descriptive generalizations about Title I projects in California, it is only the most efficient projects which are of interest for finding the true relationship of output to differing factor combination." Since, however, few schools seem very close to the efficiency frontier (in fact, they could not be anything else, in light of the apparent contradiction between the isoquants and the regressions) it seems to me apparent

that most schools can ignore the isoquant advice, unless they plan on becoming models of efficiency at the same time they are making factor substitutions.

4. Though Kiesling never does give any specific policy advice, a school administrator might still be inclined to start adding specialists to his staff—not understanding entirely that even after several years of collective effort on the part of economists, production-function studies might be best classed as simply in "the exploratory stage." But if the administrator thought for a moment, he would realize there are several very good reasons to reject the implicit policy recommendation, even if he was not bothered by the technical problems of the analysis. First, going back to my remarks at the beginning of this comment, he could argue that he simply is not the sort of administrator that works well with specialists. Second, he could say that it was the in-the-know administrators who hired a lot of specialists, realizing that was the hottest approach, the newest conventional wisdom. Such administrators are sufficiently more clever and energetic than others so that their programs would have done better in any event. Third, those programs that hire more reading specialists simply emphasized the overall goal of reading more than he tends to do, even though all those in the sample answered the survey's multiple choice question about goals the same way: that is, they all ranked improving reading scores as the first priority. The ordinal answer fails to distinguish between those administrators who saw that goal as just barely more important than some other things, and the administrators who were bent on pursuing that goal almost exclusively. The substantially different emphasis on specialists among programs suggests different degrees of emphasis; and those that hired more specialists probably tended to turn all their efforts more strongly in the direction of improving reading test scores, to the possible detriment of other educational values.

5. Though Kiesling seems to regard the programs he deals with as especially effective, and apologizes for the possible bias that this might introduce, and though he feels that the detailed survey he undertook results in more sophisticated answers, the calculated general level of success attained by these programs seems about the same as that measured earlier with a much cruder methodology. Note that the .87 of the month's gain per month of instruction highlighted in the "summary of the findings" is not the figure we are looking for. That is not a number that describes the net impact of the programs, but one which only verifies that individuals in these programs are still advancing their reading skills less rapidly than the national norm. The figures that are more relevant are the cost-effectiveness ones, and according to the information in Table 6, they suggest that the programs, on the average, give rise to a test-score gain of about 2 per cent of a year for every additional $100 spent per student. That turns out to be very close to the outcome of the most archetypical and closely controlled of compensatory education programs in operation before Title I of the 1965 Elementary and Secondary Education Act even went into effect—the New York Higher Horizons program. Whether Higher Horizons, or the California project were "successful" programs or not can be debated, but it might be noted that the

lifetime-income-earning effect of a learning gain of that size comes out to be (by the best calculations I am aware of—my own) only about half the size of the costs involved.

All the above does not lead to some grand "main point." There are a few related things that are mildly suggested for future work in this area, however. The first is that we probably should try to rely more heavily on corroborative evidence than we presently do. Every set of observations and every known technique for uncovering "what works" in education is flawed in a number of ways; and the improvements in basic data and the means for manipulating information are not undergoing such startling improvement that earlier observations deserve to be ignored. Besides going back to see if current work jibes with past observations, it would seem especially advisable to seek corroboration of findings derived from statistical inference by seeing if the same results emerge with deliberate experimentation. The problems of deliberate experiments are imposing, but they do appear more amenable to resolution—by the exercise of scrupulous care—than are the more fundamental problems involved in statistical inference. Finally, we should, perhaps, not worry excessively if production-function work never does yield clear answers that can be adopted mechanistically by school administrators. Perhaps it is enough that we simply provide concrete illustrations of how logical thinking on the question of input mix should be introduced, letting the unique circumstances and temperaments of local administrators dictate the manner in which such rules of thought are adapted.


## Burton A. Weisbrod

University of Wisconsin, Madison


Herbert Kiesling has told us a number of things about "Education as an industry":

1. While the industry produces many "products" for many markets, the process of producing *one* product, reading achievement, for one market, the educationally disadvantaged, can be examined fruitfully.
2. Firms that produce this product for this market use varying production techniques—that is, different combinations and organizations of inputs.
3. Not all of these production techniques are (a) equally *effective* in producing outputs, nor are they (b) equally *efficient*, in the economic sense that higher-cost input combinations are sometimes used when equally productive but lower-cost inputs are available.

Kiesling sets for himself the goals of *understanding* the production function (i.e., the set of technically feasible production techniques) for producing

improved reading achievement for the disadvantaged, and then of distinguishing the efficient from the inefficient techniques.

Consider first the "output" side of the production function. No one, including Kiesling, claims that reading achievement is, or should be, the only output objective of schooling or even of the Title I program. Indeed, Kiesling also refers to (although he does not deal quantitatively with) two other outputs: one for which a frequently used ordinal measure exists, mathematics achievement; and one for which an ordinal (let alone a cardinal) scale is not generally accepted, self-esteem.[1] My point in mentioning the multiplicity of goals is to remind us to consider the likelihood and consequences of *conflicts among goals*; actions that contribute to achievement of one goal may interfere with achievement of some other goal. Thus, while Kiesling's effort to discover whether one goal—improved reading achievement—is attainable for the disadvantaged is an essential first step, it must be followed by a questioning of the assumption implicit in his analysis that the various goals (outputs) of schooling are separable. If production of reading achievement contributes external diseconomies (or economies) for the production of another school output, the investigator who fails to recognize the externality will, by such piecemeal analysis, urge an inefficiently high (low) level of inputs to reading achievement.

This possibility of goal conflict is quite likely in the type of educational production process studied by Kiesling, since that process involves not only using additional instructional and physical capital resources, but also involves using additional student time. The average of "40 minutes of individualized instruction" per student per week may well mean that the student has less class time available for work that might contribute to achievement of a goal other than reading achievement.

The presence of multiple goals or multiple outputs poses a problem for the interpretation of data on the variety of production (teaching) techniques used among schools. If schooling has multiple objectives or multiple outputs, if various groups of consumers attach different relative values to the various outputs, and if the same inputs enter the production function for more than one of the outputs, then examination of the production function for any *single* output—e.g. reading achievement—would yield biased input-output coefficients. Under such conditions we would find—and indeed we *do* find—apparently widely divergent levels of efficiency in resource use among schools. The problem, of course, is the usual identification problem: do observed differences in input combinations reflect disparate economic efficiency in pursuing a given goal—the interpretation given by Kiesling—or differences in the ends being pursued?

Turning back to the objective that Kiesling considers, the production function for reading achievement, the first problem is to define "reading achievement" operationally. The fact that the State of California has "solved" that problem by requiring the use of the Stanford Reading Test may be all that a school administration needs to know, and thus it may be all that Kiesling's body of data permits him to analyze, but it by no means resolves such questions as: (1) Insofar as students' performances on the various

widely used reading tests are not perfectly correlated, then how sensitive are our production-function estimates to the choice of a particular measure of reading output? (2) To what extent are instructors "teaching to the test"—that is, in effect, giving children the answers to the test so as to achieve high "performance" levels (at least in a single year)? In order to cope with this latter problem, Office of Economic Opportunity, in its experimentation with "performance contracting" for compensatory education of the disadvantaged, selected randomly from a set of five examinations to be administered to each school group. There is a dilemma, however, which even this procedure cannot escape. On the one hand, insofar as the various tests provide results that are strongly and positively correlated and the teachers know it, using multiple output measures is of small value—teaching to one test is teaching to all of them. On the other hand, insofar as the results are *not* highly correlated, the measured "success" of any teaching effort (producing technique) will depend on the particular measure of output (reading achievement) that is used!

Kiesling does appear to be uneasy about the particular output measure used in his study, for he refers repeatedly to the "difficulty" of the Stanford Reading Test. Unfortunately he does not tell us what he means by "difficult" although I might guess that he means that an improved performance level on *this* test is associated with a greater relative improvement in performance on some *other* reading tests. By this interpretation, the different improvement factors (*output* or achievement-added) on the various reading tests do indeed pose a problem for anyone trying to estimate the production function for the product, "reading achievement." With the product being the improvement in test score, the choice of a particular test becomes important. It would be desirable, indeed, to know how sensitive Kiesling's productivity estimates are to the choice among output (test) measures.

Turn now from output to the inputs. I stated above that we could think of "disadvantaged" students as the *market* for the Title I compensatory education programs. But at the same time, the student is also an *input* to the production process, as is frequently the case with personal services. He is an input in the sense that he must be present (physically) at the point of production and must "cooperate" (mentally) with other inputs.

Which *specific* characteristics of an individual student are important for determining the degree to which he will benefit from exposure to a particular set of instructional inputs? Kiesling hypothesizes that the important characteristics are, or at least have as their proxies, the following three: (1) being black (or a member of some other "minority"), (2) being on AFDC, (3) having moved recently or being in a group with many other students who have moved recently. Kiesling is convinced that the lack of importance that he finds for these "socioeconomic variables is attributable to the fact that undoubtedly . . . they are too crude." Perhaps he is right. But perhaps socioeconomic class is simply unimportant for explaining variance in reading achievement—once a "more fundamental" variable is taken into account. The more fundamental variable to which I am referring is the achievement level from which the student *began* his participation in the remedial pro-

gram. Indeed, the variable, "beginning-score," was one of the most statistically significant variables considered by Kiesling. (While the significance level was high, the actual quantitative importance was more modest; for each ten months in which the student was below the reading norm for his grade, participation in the group contributed approximately an additional one-third month to the increase in reading performance during the remedial program.)

The significance of the beginning-score variable—as an indicator of student achievement at the onset of the program—deserves more attention than Kiesling gave to it. Whereas he considers it only as a variable entering additively with other variables, there is a distinct possibility that such variables interact with, and hence, condition, the effect of other input variables, such as the type of instructional approach used. Indeed, my colleagues, W. Lee Hansen and Allen Kelley, and I, in a 1970 article,[2] pointed to evidence of such an interaction effect, although involving higher education. We noted that a given instructional technique might vary considerably in effectiveness among students of differential initial achievement levels, and, referring to some current experimental work by Kelley, we noted that one new instructional technique was consistently most effective (in terms of enhanced performance on a standardized achievement test) for students at the highest initial levels of achievement.

But the inputs which Kiesling appears to have most interest in, and most faith in, are the planning, coordination, leadership, and teamwork variables. His faith does not appear to be shaken by the statistically insignificant coefficients he finds for variables reflecting time spent in planning and in-service training, percentage of key people who are involved in planning, and the degree of teamwork. I suggest that Kiesling's faith in these variables is misplaced, not because they are necessarily unimportant—he is probably right in blaming "poor data"—but because they are not now *instrumental variables*. Coordination, leadership, and teamwork are not variables that can easily be added to or subtracted from a production process; we know so little about how to produce these important inputs. I believe that Kiesling is implicitly recognizing this when he admits that "teamwork is an extremely difficult thing to quantify," and that "it is also difficult to correctly represent the degree of informal meeting and discussing that occurs" among regular teachers, specialists, and leaders in some remedial programs.

To say that these variables are difficult to quantify is to say, in effect, that we do not know how to vary them. Thus, even if Kiesling is right in his faith in something called "leadership, coordination, and teamwork," we cannot depend on those variables for successful remedial programs for the educationally disadvantaged until we discover how to produce them.

A final word concerning the *dating* of variables in the production function. Kiesling implicitly set out to estimate a production function for remedial reading in a manner analogous to the production function for corn or wheat: inputs are applied during a period and a single output (crop) comes at the end of the period. As a first approximation this is an attractive approach, but it is probably a considerable simplification of reality. Learning is probably

more of an *investment* process in which "outputs" from one year of schooling continue to flow for perhaps many years, thereby affecting ability to learn in subsequent school years. That is, just as inputs used in growing bananas (unlike corn and wheat) will contribute not only to the output of bananas in that year but will contribute to banana production in subsequent years, so remedial schooling inputs in one year may have effects beyond that year. Indeed, we hope that is the case!

Unfortunately, we cannot be certain that such future "output" effects will be positive. They can be negative. Analyses of OEO Head Start Programs, for example, has disclosed that output (reading level achievement) did increase during the period of the program, but one or two years later most or all of the added achievement had apparently vanished; there was no longer a difference between the achievement level of students who had been part of the Head Start Program and the achievement of a control group of students who had not. There are, of course, various interpretations of these findings; the point is only that "outputs" can be quite different, depending on the date at which they are examined.

We—economists or educators—know so little about the production function for any type of educational output for any market segment that empirical studies such as Kiesling's can be usefully multiplied manyfold even before diminishing returns set in—and, of course, we should not stop there. As this work proceeds, though, we need to place greater emphasis, I believe, on building models in which (1) schooling is treated as an investment process yielding multi-period returns in terms of learning; in which (2) interactions among inputs are considered; in which (3) inputs that can be subject to control are emphasized; and eventually (4) on building models which account for conflicts and complementarities among multiple goals.

## NOTES

1. The much-praised "Sesame Street" television program has come under attack for its alleged failure to develop students' self-esteem. See, for example, *Newsweek*, May 24, 1971, p. 52.
2. W. Lee Hansen, Allen C. Kelley, and Burton A. Weisbrod, "Economic Efficiency and the Distribution of Benefits from College Instruction," *American Economic Review* 60 (May 1970): 364–369.