DOCUMENT RESUME

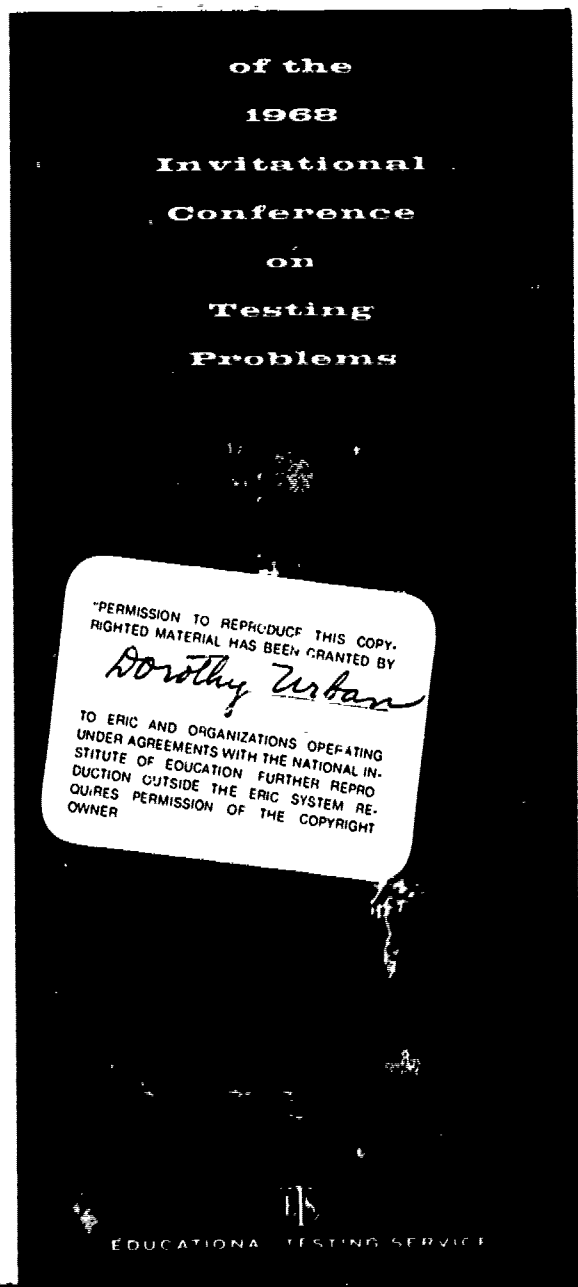
ED 080 542 TM 003 047

TITLE	Invitational Conference on Testing Problems (New York, November 2, 1968).
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	2 Nov 68
NOTE	149p.
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	Comparative Analysis; *Conference Reports; Cost Effectiveness; Disadvantaged Youth; Employment Practices; *Evaluation Methods; *Program Evaluation; *Socially Disadvantaged; Teacher Education; Teacher Educator Education; *Testing Problems

ABSTRACT

The 1968 Invitational Conference on Testing Problems dealt with educational evaluation and the problems of the socially disadvantaged. Papers presented in Session I, Educational Evaluation--Various Levels and Aspects, were: (1) "The Comparative Field Experiment: An Illustration from High school Biology" by Richard C. Anderson; (2) "Evaluation of Teacher Training in a Title III Center" by Ethna R. Reid; (3) "Evaluating a National Program: The Training of Teachers of Teachers" by Bertram B. Masia and P. David Mitchell; and (4) "Cost-Benefit Analysis and the Evaluation of Educational Systems" by J. Alan Thomas. The luncheon address was "A Customer Counsels the Testers" by Sidney P. Marland, Jr. Papers given at Session II, The Socially Disadvantaged, were: (1) "Nonschool Variables in the Education of Disadvantaged Children" by Edmund W. Gordon; and (2) "Issues and Strategies in Employment of the Disadvantaged" by Albert P. Maslow. One paper presented at the second session, "Meaning, Thinking, and Reading" by Marie Hughes, was not completed in time to be included in the proceedings. (KM)

PROCEEDINGS



Library of Congress Catalog Number: 47-11220
Printed in the United States of America

Invitational Conference on Testing Problems

November 2, 1968
Hotel Roosevelt
New York City

J. THOMAS HASTINGS
Chairman

EDUCATIONAL TESTING SERVICE
Princeton, New Jersey
Berkeley, California
Evanston, Illinois

Foreword

Educational evaluation and the problems of the socially disadvantaged are concerns of the utmost importance and relevance to educators today and as such provided timely themes for the 1968 Invitational Conference. The morning session was devoted to an exploration of various levels and aspects of evaluation. The first speaker described a field experiment designed to compare the effectiveness of two high school courses in population genetics. The second and third speakers were concerned with the evaluation of programs of a broader scope: a Title III center for reading instruction and a national program for the training of teachers of teachers. The final paper of the morning dealt with an approach to analyzing school systems by means of a study of costs and benefits.

The afternoon session, on the theme of the socially disadvantaged, explored this subject from the point of view of a teacher, a teacher of teachers, and a Civil Service psychologist. They spoke on the problems of reaching children of minority groups, the function of the school in the ghetto, and the need to expand job opportunities for the nation's 500,000 hard-core unemployed.

Complementing the themes of the two sessions was Dr. Sidney Marland's luncheon address, "A Customer Counsels the Testers," a provocative and constructive critique of the state of the art of academic testing today.

I should like to express our thanks to Dr. J. Thomas Hastings who, as chairman, created the program and to the speakers whose papers appear in these *Proceedings*.

Henry Chauncey
PRESIDENT

Preface

The Invitational Conference on Testing Problems has long been a major convocation among the various annual meetings of those who are concerned with educational measurement. Over the years, the main focus of the papers delivered has been upon improved measurement and interpretation of measurement of student outcomes in formal schooling. Professionals and scholars, over these same years, have been asked to help with improved evaluation of educational endeavor. Their efforts have combined more accurate and more precise measurement of student outcomes with careful design of data collection and analysis of results. The main concern has continued to be on results of instruction—both formal and informal—in connection with individual students, although group results are studied in looking at "programs" as a summation of individuals. In past *Proceedings* one can find a considerable amount of material on intelligence and aptitude tests, as opposed to direct instructional outcomes, but the emphasis here has been upon prediction and interpretation of the results of instructional treatment.

In the last few years noticeable increases in at least two elements of education have tended to bring about an alteration in the demands of the educational system for evaluation. Educational measurement people are still being asked to do educational evaluation, but the requirements have changed. The changed elements are, first, the development of large national programs such as Title III, Title I, the Triple-T Program, and Head Start. The second element altered is that of the school's relationship to the total community. Two years ago, at the 1966 Invitational Conference, the luncheon speaker, Francis Keppel, said, "There was a time, not too long ago, when education was thought of, more often than not, as its own little universe, as a

thing apart from the rest of society. That is no longer nearly so true. Education has become more and more involved with the rest of society, with government, with industry, with all manner of agencies and institutions. The problems that beset all of us—urbanization, the population explosion, automation, communications, and so on—are also education's problems, both in the sense that they affect education and in the sense that education is helping to solve them." If such change was noticeable then, it is downright obvious now. These changed elements have put new demands on education which have been transmitted to the people in educational measurement whom the establishment is accustomed to asking for help in evaluation. This, in turn, means that those in educational measurement, if they are to be engaged in such evaluation, must become used to talking in terms of the new kinds of evaluation tasks.

The morning session grew out of these concerns. I received help from a large number of people in terms of suggested names or topics for the morning session, which was to be devoted to the larger theme of evaluation. Thus, the first paper, by Richard C. Anderson, deals with the evaluation of a small part of an instructional program and deals with it in an experimental fashion. The focus, however, is not solely on testing of student outcomes, although such measurement plays a large part. The presentation by Ethna R. Reid focuses upon some of the problems and issues involved in evaluating a large Title III center. Again, although testing and measurement play a part, there is far less focus on student outcomes than even in the first paper. The paper by Bertram B. Masia and P. David Mitchell picks yet a larger system with a different focus. They speak to the variables and data sources in evaluating a large national program—the Triple-T. The last paper in the morning session, by J. Alan Thomas, is devoted to a discussion of some of the problems and theories involved in doing a cost-benefit analysis of a school system. Some of the things said in each of the four papers are somewhat removed from the usual concerns of the educational measurement person who is used to dealing with student outcomes and their prediction. For that very reason, it is hoped they will help such specialists become aware of the relations between their measurement of behavior in individual students and the larger problem of evaluation of educational efforts as they have developed today.

The general topic of the afternoon session is not far removed from that of the morning session. It aims at elucidating the connection

between schools as we have understood them and the larger society through the urgent business of working with those segments of society which we have referred to as "disadvantaged." Edmund W. Gordon's presentation sets forth some of the educational variables which affect instruction but which are not contained within the school system when we consider the school as "its own little universe." The final paper of the afternoon, by Albert P. Maslow discusses an area which is extremely important in the total school-societal endeavor in terms of interaction in both directions. It enunciates some special problems and variables in the *employment* of the so-called disadvantaged. Although each of these afternoon papers touches upon testing problems connected with individuals, they certainly are not focused upon that operation as an end in itself.

Unfortunately, the first paper of the afternoon session, "Meaning, Thinking, and Reading," by Marie Hughes of the University of Arizona, is not included in these *Proceedings*. Because of illness and the press of many commitments, Dr. Hughes was not able to finish her paper to her satisfaction before the final deadline. This is truly an unfortunate loss for those who were not able to attend the conference.

I have been honored by being invited by the three former chairmen to be chairman of the 1968 Conference. The choice of topics and speakers was mine with much help and advice from others, but the operation of the conference was possible only through the efforts and good judgment of many people at ETS—particularly Anna Dragositz. Henry Chauncey's well-known graciousness and support make the conferences great and the chairing task a delight.

However, the speakers really make a particular conference. My thanks to these speakers for excellent presentations. I feel certain that your careful study of the papers will add to your understanding of the complexity of the variables and the variety of the expertise needed in looking at educational evaluation today.

J. Thomas Hastings
CHAIRMAN

Contents

iii Foreword by Henry Chauncey

v Preface by J. Thomas Hastings

Session I: Educational Evaluation: Various Levels and Aspects

3 The Comparative Field Experiment: An Illustration from High School Biology, Richard C. Anderson, University of Illinois

31 Evaluation of Teacher Training in a Title III Center, Ethna R. Reid, Exemplary Center for Reading Instruction, Granite School District, Salt Lake City, Utah

72 Evaluating a National Program: The Training of Teachers of Teachers, Bertram B. Masia and P. David Mitchell, Case Western Reserve University

89 Cost-benefit Analysis and the Evaluation of Educational Systems, J. Alan Thomas, University of Chicago

Luncheon Address

101 A Customer Counsels the Testers, Sidney P. Marland, Jr., Institute for Educational Development

Session II: The Socially Disadvantaged

115 Nonschool Variables in the Education of Disadvantaged Children, Edmund W. Gordon, Teachers College, Columbia University

123 Issues and Strategies in Employment of the Disadvantaged, Albert P. Maslow, U.S. Civil Service Commission

Session I

Theme:
Educational Evaluation: Various Levels and Aspects

The Comparative Field Experiment: An Illustration from High School Biology*

RICHARD C. ANDERSON
University of Illinois

A time-honored, but not otherwise honored, form of educational research is the experiment comparing different methods of instruction. Throughout the years there have been numerous comparisons of televised lectures with live lectures, discovery teaching with expository teaching, learned-centered instruction with teacher-centered instruction, programmed instruction with textbook instruction, and so on. The lessons employed in these experiments were of little interest for their own sake. They were merely the vehicles for evaluating a method of instruction it was *assumed* would apply broadly to many lessons. Today I dare say there is general agreement that this was a poor assumption (9, 12). Nonetheless, I shall argue that the comparative experiment, recast for a different purpose, deserves a share of the time and energy of the educational research community.

The argument begins like this: We have little power to predict the kind of instruction that will work best with learners. There are no methods of instruction that have proved consistently better than, or even consistently different from, other methods. There are no procedural features of lessons that are invariably associated with greater student achievement. Neither small steps, nor active responding, nor immediate feedback, nor a warm classroom climate, nor a sequence from concrete to abstract, nor the provision for self-direction and

*The author is deeply indebted to Gerald Faust, John Guthrie, and Veronica Drantz, who helped prepare the instruction; to Gerald Faust, Marianne Roderick, and Phillip Zediker, who assisted in the collection and analysis of data, and to Robert Stake, who criticized a draft of this paper. The research described herein was supported in part by a grant from the National Science Foundation.

self-pacing, nor multi-media stimulus bombardment—singly the aggregate—guarantee successful instruction.

This is a bleak picture, but I think not overdrawn. To be sure, we do know some things. However, there are more things about which we are uncertain. There are many anomalies. It is, in my judgement, a fair appraisal to say that right now, today, we cannot confidently predict the effectiveness of a lesson, given a description of its philosophy, the style, the methods, and the procedures of instruction.

If this is a fair appraisal, then consider the following question: How should the money of funding agencies and the time and energy of the educational research community be allocated to maximize the effectiveness of instruction today and in the future? One answer is to invest in basic educational and behavioral science research. I am very sympathetic with this view. We should, however, be realistic about the impact basic research can have on instructional practice.

There is currently a high "degree of empiricism" (8) in the behavioral sciences. This is doubly true of the applied sciences that look to behavioral science for inspiration. Basic educational research should gradually increase our capacity to specify in advance of tryout the edures and arrangements of instructional material which are qui .tain to facilitate student learning. Gradual progress can be counted on. But it would be unrealistic to expect that we shall ever be able to predict effective configurations of instruction with any more than modest confidence. I have do doubt that instructional development will always be based in part on rules of thumb. And I have no doubt that a considerable amount of trial and error will always be necessary to *guarantee* successful instruction.

Thus far I have argued that we have little power to predict the features of instruction that will maximize student learning, and that basic research can be expected to add only modest increments to our capacity to predict. There is one thing we can do right now. We can distinguish in tryouts or field tests units of instruction which teach well and units of instruction which teach poorly. The rational thing to do is to make use of this capability to improve instruction. The contention is that lessons should be evaluated in terms of the results they produce with students and that every step in the development of instructional materials should be guided by student performance. It is not possible on the basis of existing knowledge to warrant methods of instruction, but it is possible to warrant particular lessons, units, or curricula.

Satisfactory lessons can be developed using systematic trial and error. The process consists of the designation of objectives, preparation of instructional materials that hopefully meet these objectives, and then tryouts of the materials with people of the type for whom the instruction is intended. On the basis of successive tryouts, revisions are made in the materials. The process of tryout and revision continues until objectives are being reached or the decision is made that it is impossible to reach the objectives within the limits of available time and resources.

The Role of the Field Test

When the pilot test data indicate students are reaching objectives, it is time for a field test of the instructional materials. Better to say that a total instructional package is to be field tested. An "instructional package" includes not only materials put directly into the hands of students and scripts or guides to the teacher indicating how to conduct discussions, laboratories, problem-solving episodes and so on, but it also may include teacher manuals, provision for teacher workshops, and the supervision of instruction. One purpose of the field test is to determine whether the instructional package will perform successfully under various conditions of use. The pilot test may or may not have involved the full range of students who will use the materials. The pilot tests were presumably completed or supervised by someone who was enthusiastic about the project and knowledgeable about the use of the materials. What will happen if the materials are placed in the hands of teachers who are indifferent or even hostile? Must the materials be used in a certain way or will they be reasonably successful under a variety of conditions? If the curriculum must be used in a certain way, is there provision for teacher manuals or workshops? And are the manuals or workshops successful in getting teachers to behave in the intended way? These are some of the questions which can be answered in a field test.

To use the terms introduced by Scriven (13), the purpose of pilot tests is formative evaluation, to locate weaknesses in student understanding or performance so that editors, writers, or teachers can revise and presumably improve instructional materials and procedures.

Providing feedback to the developers of the instruction is only a

secondary purpose of the field test. The chief purpose is summative evaluation. Data are gathered in order to help potential consumers—education agencies, administrators, teachers, students—make the decision to use or not to use the instructional package.

Some people who argue for empirical validation of instructional materials seem to take the position that effectiveness in modifying student behavior is the sole criterion for judging instruction. Let me emphasize that this is not my position. Lessons, units, and curricula should be judged in terms of the extent to which they reach their goals, but this cannot be the only criterion. Other criteria include the cost of the instructional sequence in terms of student and teacher time, the acceptability of the sequence to students and teachers, and any side effects (14). The accuracy, up-to-dateness, and elegance of the subject matter has been the important criterion for the prominent curriculum reform projects. A most important criterion is the worthiness of the goals the instruction aims to reach. As Scriven (13) has noted, "it is obvious that if the goals aren't worth achieving then it is uninteresting how well they were achieved." A complimentary assertion is also true: No matter how worthy the goals, a lesson cannot be valued highly if it is ineffective in reaching these goals. Effectiveness should be neither overrated nor underrated as a criterion for judging instruction.

Sometimes the field test of an instructional package should be a comparative experiment. This conclusion is inescapable if the function of the field test is to guide consumer decisions. There are several alternative sets of instructional materials for most areas taught in school. To the extent that the objectives and the coverage of different sets of materials overlap, it is entirely reasonable for the practical decision maker to ask which is most effective.

Cronbach (9) and Scriven (13) have taken opposing sides as to the value of comparative experiments. With one qualification, I agree with Scriven: Comparative experiments do have a valuable function. But Scriven seems to envision massive Consumer Union-type comparisons of the curricula within every subject-matter area. To this plan, Cronbach has rightly objected that most comparisons probably would show no differences of statistical or practical significance.

Comparative experiments are expensive. They cannot be run indiscriminately. One criterion for deciding whether a comparative experiment should be run is this: There must be considerable confidence that one of the instructional packages is in fact more effective than the

other. The playing of hunches has its place in basic research. But this strategy cannot be encouraged in comparative educational research. From the point of view of the person undertaking a comparative experiment, it should simply be a demonstration that one of the instructional packages is superior.

In a comparative experiment, the no-difference result has very low social utility. If you firmly reject the notion that a comparative experiment can show the general worth of a method of instruction and accept the concept that the main justification for the comparative experiment must be to determine which of two or more particular lessons (instructional packages) is the most effective, then obviously it doesn't make any sense to compare lessons on the mere chance that one of them might be better; unless, perhaps, you believe there are many great lessons, unrecognized, lying around waiting to be discovered. Perhaps there is some value to finding that a widely touted curricular innovation is no more effective than other instruction. In general, though, inconclusive comparative experiments cannot facilitate consumer decisions. Consequently, there has been error in judgment when a comparative experiment shows no difference. Time and money, which should have been invested in instructional development and formative evaluation, have been wasted on a premature comparison.

It may be argued that research cannot be justified merely to demonstrate what is confidently believed to be true. The counter-argument is based on the thesis developed earlier in this paper. There are no a priori tests that reliably predict lesson effectiveness and no experts whose acumen in judging lesson effectiveness has been established. In short, there are no acceptable grounds for claims about effectiveness of a lesson except for data which actually demonstrate effectiveness.

The Need for Relative Standards

The position that lessons should be evaluated in terms of absolute standards of effectiveness is widely voiced. Indeed, this is the position I take with respect to pilot tests of lessons. When it comes to field tests of lessons, there are reasons to be cautious about relying entirely upon absolute standards. First of all, unlike other practical fields ranging from agriculture to the automobile industry, education has no consensual standards of performance.

Suppose that educators could agree on some general standard, let us say the famous 90-90 standard propounded by the Air Force Training Command under the leadership of Colonel Gabriel Ofiesh. What would it mean if students averaged 90 percent on a criterion measure? Obviously it would not mean that the students had mastered 90 percent of all there is to know about a topic. What it would mean is that they had learned 90 percent of what someone chose to teach and measure. Herein lies the problem. Despite recent advances in clarifying educational objectives, there can still be considerable variance in the intended or implicit level of sophistication with which a concept is developed, given ostensibly the same goals. A further problem is that the level of performance is a function of measurement technique, including such factors as, for instance, the attractiveness of distractors in multiple-choice items. Finally, the fact that an instructional package meets a certain standard of effectiveness does not preclude the possibility that a competing package will exceed the standard in less time at lower cost. The conclusion is that relative standards, and therefore comparative experiments, are necessary to gauge the effectiveness of instructional materials.

Do not mistake me. I believe that the notion of absolute standards of effectiveness is sound in principle. I hope that it will be possible to augment the logic and the techniques for formulating absolute standards. For the indefinite future, in consideration of our fallibility in defining absolute standards and measuring performance with respect to them, absolute standards should be supplemented with relative ones. At the present time, the only dependable way to determine which of two lessons is most effective is to directly compare them.

There is a clear role for the comparative experiment when several lessons (units, curricula) have substantially the same goals. When this is the case, the best lesson is the most effective lesson (assuming other factors such as cost are comparable). The consumer can concentrate mainly on the data from a comparative experiment when choosing among instructional packages. Moreover—and this is one of the reasons I am advocating comparative experiments—the competition to produce better lessons will itself promote more effective instruction.

It is not so clear to me that there is a role for the comparative experiment when the goals of several instructional packages are different. Another open question is: Who should run comparative experiments, the developers of new instructional packages or independent evaluators? There are also questions regarding the proper design

and execution of comparative experiments. Rather than address myself to such questions in general terms, I shall now attempt to practice what I preach. The rest of this paper describes a comparative field experiment completed to demonstrate the effectiveness of some new instructional materials.

THE FIELD TEST OF A PROGRAM IN POPULATION GENETICS

Development of Experimental Materials

Under the sponsorship of the Biological Sciences Curriculum Study, a self-instructional program in population genetics was prepared for use in high school biology (10). The program was developed using the procedures briefly outlined earlier. As a first step, objectives were defined. In this regard, the treatment of population genetics in the Biological Sciences Curriculum Study textbooks served as guides. Next, a trial version of a portion of the program was prepared. This segment was tried with a series of individual high school students with one of the authors of the program monitoring each student's performance. After every few students, revisions were made. The remaining segments of the program were developed in the same fashion. Eventually the entire program was tested with small groups of students. Once again revisions were made. Throughout the development of the program, a lengthy criterion-referenced test—consisting largely of open-ended, constructed response items, problems to be solved, concepts and principles to be defined and illustrated—was used. Pilot subjects spent nearly as much time completing the criterion test as they did receiving instruction. Generally, a segment of the program was judged satisfactory when all pilot subjects scored 90 percent or better on criterion test items keyed to that segment. The edition of the program used in the experiment consisted of about 14,000 words, exclusive of equations and graphs, in 234 frames.

Comparison Instruction

The program on population genetics was compared with the treatment

of population genetics in the textbook, written by the Biological Sciences Curriculum Study, entitled *Biological Science: An Inquiry Into Life* (5), unofficially known as the "BSCS yellow version." This text contains about 7,900 words dealing directly with population genetics. The material in the textbook was supplemented by laboratory exercises also prepared by the Biological Sciences Curriculum Study, and the instruction was of course mediated by a high school biology teacher. It would be wrong to label the comparison instructional treatment as "conventional instruction." This material had been prepared by a team of biologists and biology teachers. The textbook had undergone a major revision based in part on the systematically collected opinions of many teachers around the country who used the experimental edition. There is a prima facie case that there is no better instruction in population genetics for high school students than the BSCS yellow version and auxiliary materials.

Design of the Study

Approximately 750 high school biology students taught by 9 teachers in 30 classrooms in 2 suburban high schools participated in the experiment. Each of the nine teachers taught from two to four classes. Classrooms were randomly assigned to receive the program with the restriction that, when possible, half of each teacher's classes received the program and half did not. Two forms of the achievement test were developed. Within each classroom a random half of the students received one form as a pretest and the other as a posttest. The assignment was reversed for the remaining half of the subjects. This design provided baseline data and furnished information on a relatively large number of items with a relatively small investment of student time, but it avoided the specific priming effect that can occur when students repeat exactly the same test.

Procedure

The participating teachers agreed to give the pretest and the posttest on certain dates. In between, the teachers agreed to use the program with designated classes and not with others. The teachers were told: "Please use the program according to your best professional judg-

ment." A member of the project team briefly talked with the teachers about procedures for giving tests and recording the data but he made no attempt to sell the program or to recommend how it should be used.

The rather unstructured directions to teachers can be understood in terms of the question the experiment was intended to answer, which was: Is the self-instructional program on population genetics a valuable supplement to other BSCS materials on this topic? We wanted to see whether the program was of value under conditions of ordinary use because under such conditions teachers will employ materials as they see fit.

In retrospect, there is no doubt that better overall results would have been obtained with the program if a teacher's manual had been furnished as part of the instructional package studied in the experiment. However, at the time, and I still think wisely so, we decided not to provide a manual. A manual is functional only to the extent that teachers follow the advice it contains, and our experience has indicated that teachers often do not read the manuals which accompany instructional materials. It was judged important to discover how failsafe the materials are under adverse conditions. A teacher's manual is currently being prepared, and the results of the experiment are being used to convince teachers to follow the recommendations it contains.

Stake (14) has convincingly argued that a sound evaluation of an instructional treatment must include a full description of the manner in which the treatment was implemented. Such description was especially important in this case because teachers were permitted a great deal of latitude. Teachers completed a log for both program and no-program classes describing all classroom activities between the pretest and the posttest and the number of minutes devoted to each activity. All laboratories, exercises, and reading assignments were specifically detailed. Teachers also completed a questionnaire, containing both open-ended and forced-choice questions, that inquired about the attitudes of the teacher and his students toward the program, techniques for using the program and relating it to other class work, and strengths and weaknesses of the program. Students answered a questionnaire covering similar topics.

RESULTS

Overall Analysis of Achievement Gains

Since whole classrooms were randomly assigned to program and no-program conditions, the unit of observation was the classroom. The datum for the overall analysis of variance was mean classroom achievement gain. Included in the classroom means were the scores of all students who were present for the pretest and the posttest and who received the appropriate forms of the tests. A number of individual students and one entire classroom were discarded because they failed to meet these criteria.

An unweighted means analysis of variance was performed with program or no program and school as the factors. Only the difference between program and no program was significant [$F(1, 25) = 20.59$, $p < .01$)]. An ω^2 of .39 was obtained. In other words, 39 percent of the variance in classroom achievement gains was accounted for by treatment. The actual gains were 4.62 items under the program condition and 3.01 items under the no-program condition. Relatively speaking, this means that those who received the program gained 53 percent more than those who did not; however, the absolute difference was not as large as we had confidently expected. The reasons for the failure to find a larger absolute difference will become apparent later.

Achievement as a Function of Source of Test Items

A criterion-referenced test featuring constructed-response items was used in pilot tests of the program whereas a norm-referenced multiple-choice test was employed in the field experiment. There were two reasons for this switch. The first was a simple matter of expediency. We didn't dare ask cooperating schools for enough time to give a longer test. The second, and more important, reason was to insure the credibility of the results in the eyes of the consumer whose decisions the experiment was intended to influence. In the present case the Biological Sciences Curriculum Study is the immediate consumer. This organization has spent a good deal of time and money to develop unit achievement tests which, among other topics, cover population genetics. Since the program was designed to reach the same objectives

as the other BSCS materials in this area, it was hard to make a convincing argument for not using the test items which biologists and teachers of biology believe validly measure student performance in relation to these goals. Criterion-referenced tests are the only ones which make sense for the evaluation of instruction, but in this case it was essential to use the BSCS norm-referenced items to avoid the suspicion that the program shows to advantage only on especially tailored items.

Twenty-four items in the BSCS unit tests which dealt with population genetics were included in the achievement test used in the experiment. It should be emphasized that a difficulty level of near 50 percent *after instruction* has been one of the criteria for including items in BSCS unit tests. Twenty additional multiple-choice items were created to complete a table of specifications. Figure 1 pictures achievement gains for the program and no-program conditions as a function of the source of the test items.*

Achievement as a Function of Content Area

The objectives of the program can be classified into three topical content areas; first, the calculation of gene proportions, given sample data; second, the logic of the Hardy-Weinberg principle, and third, factors that cause gene pools to change (mutation, adaptation-selection, differential migration, random genetic drift, nonrandom mating, isolation). The program compromises on a fourth content area. Mastery of basic Mendelian genetics is essential to learn population genetics. The student is presumed to have covered basic genetics before he begins the program. However, rather than depend upon the adequacy of previous instruction, basic genetics is reviewed at the beginning of the program. Figure 2 shows gains in achievement in each of the four content areas.

Achievement as a Function of the Type of Item

One of the potential weaknesses in the technique of revising instruction until performance on a criterion test reaches a satisfactory level

*Percent gain is simply actual gain divided by total possible score.

Figure 1

Achievement Gains According to the Source of the Items

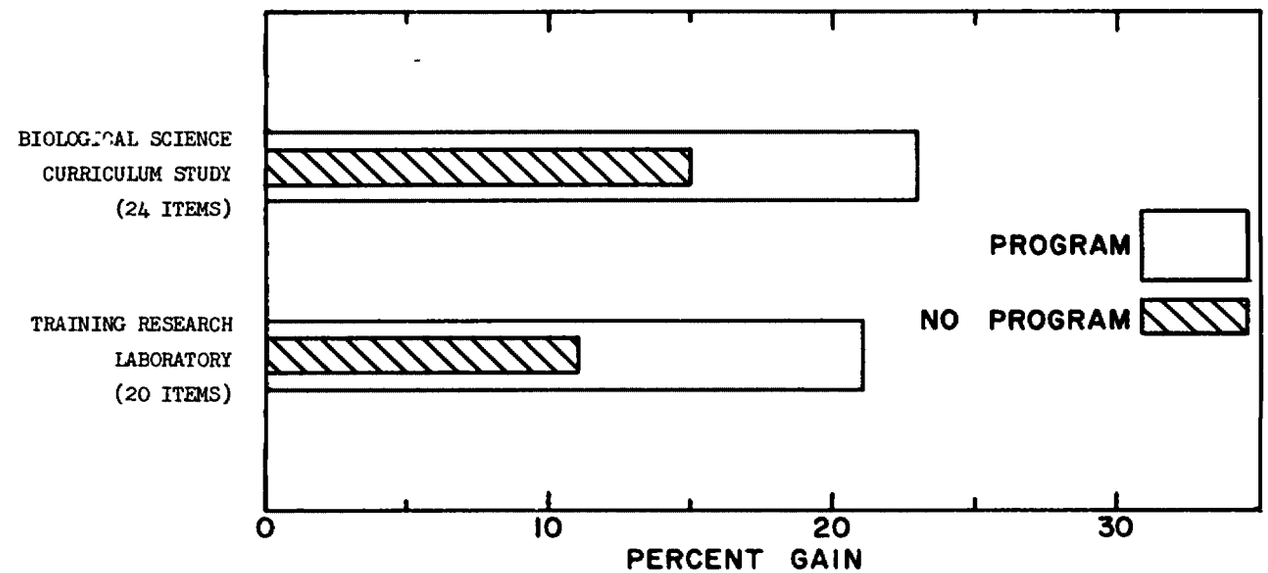
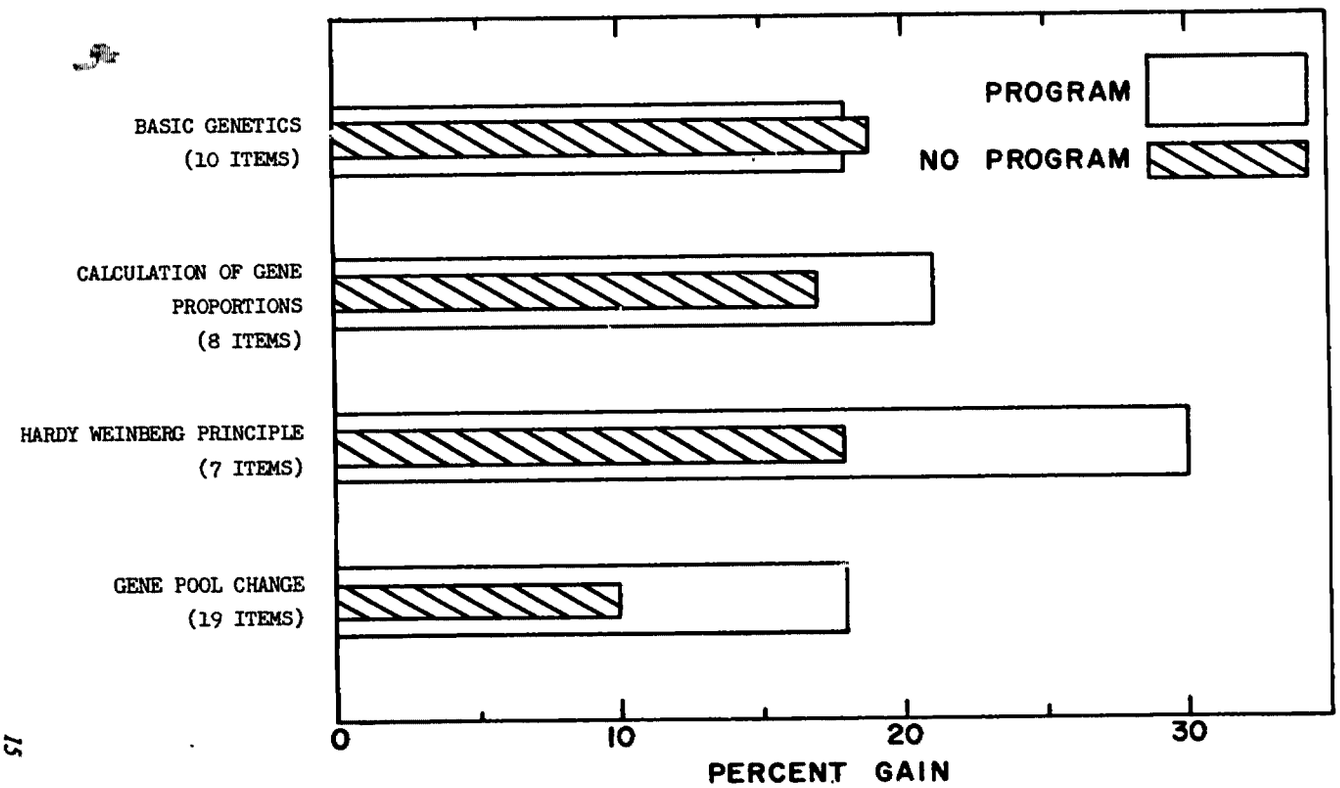


Figure 2

Achievement Gains According to Content Area



is that this technique may lead to a trivial sort of teaching to the test. What can happen is this: Performance is poor on a test item, so the author or programmer includes statements within the instruction which give the answer to the question; or perhaps he asks the question during the course of instruction in a context in which the student cannot fail to answer correctly. Performance on the offending test item is bound to improve. But the question of what has been learned must be asked. It may well be that students have only learned to repeat or recognize a string of words. By definition a person "understands" a concept or principle when he is able to deal in an appropriate manner with any member of the universe of instances covered by the concept or principle.* If a test item involves an example which appeared during instruction, it may entail no more than verbatim recall or recognition. However, if the student is able to correctly answer items involving instances from the universe which are *different* from any of the illustrations presented in the lesson, then it is proper to infer that the person understands the concept. The instances contained in test items can be scaled in terms of the degree of similarity to instructional examples. A person who can answer questions containing instances slightly different from those used in the lesson can be said to have some understanding of the concept or principle, whereas the person who can answer items containing *very different* examples has a "deep" and a "broad" understanding.

Concepts can be defined in generic terms, and principles can be stated in abstract, general language. If a test item substantially repeats the language of instruction, once again only verbatim recognition is required to answer correctly. However, if the student can deal appropriately with expressions of the concept or principle which are worded differently from, but are semantically equivalent to, instructional statements, some level of understanding is implied.

A capacity to synthesize is implied when the student can correctly answer a test item which entails applying concepts or principles treated at widely scattered intervals during instruction. Alternatively, these items can sometimes be answered correctly if the student extrapolates or draws an inference from statements made in one place during instruction.

Using the distinctions just outlined, a content analysis of the in-

*Note that a curriculum may aim to produce appropriate performance with only a subset of the universe

struction and the test items was completed. Each item was assigned to one of five categories on the basis of the relationship between the wording or example in the ite , and the language and the examples contained in the program. Neither the textbook, the laboratory exercises, nor, of course, the oral discourse of the teachers were considered. I must warn you that I cannot vouch for the reliability of the assignn.:nt of items to categories. This is to be regarded as a rough, beginning attempt to operationalize the ideas which educators have regarded as important since the work of Bloom (6) and his associates.* Figure 3 contains achievement gains under the program and no-program conditions as a function of the type of item. Only one item was judged to measure verbatim recognition, so this category is not included in the figure.

On a previous page, I stated that test items containing instances very different from the examples used during instruction could give evidence of "deep" and "broad" understanding. Bear in mind that such items may test beyond the objectives of a particular curriculum. While they probably should be included in achievement tests to probe the limits of understanding, caution is indicated in judging instructional packages in terms of effectiveness with items which contain very different examples. To put the matter another way, this type of item measures "remote" transfer of training, an unreliable phenomenon which no treatment can be counted on to produce.

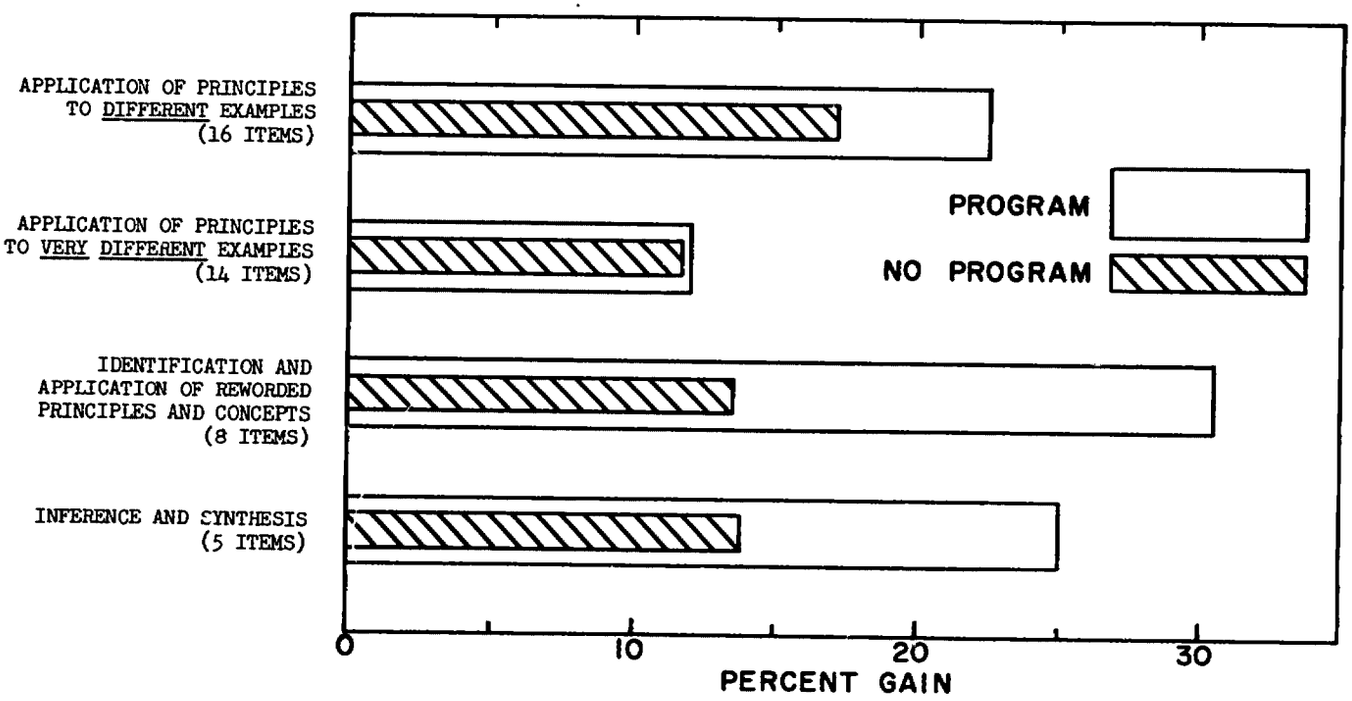
Achievement as a Function of the Teacher

There was enormous variation in the ways teachers used the program. Some teachers did not allow any class time for students to study the program while, at the opposite extreme, there were teachers who used the program and only the program to teach population genetics. Table 1 gives the number of classroom minutes devoted to different activities between the pretest and posttest. These numbers are based on the teacher logs. We proposed to the schools a two-week interval between the pretest and the posttest. Teachers at one school said "Fine." Those at the other school said "We can't possibly teach this material in less than a month," so they got a month. All of the teachers in School B reported spending the same amount of class time on population genetics

*See Anderson (in press) and Chapter 7 in Anderson and Faust (in press) for a further explication of these ideas.

Figure 3

Achievement Gains as a Function of Type of Item



whether or not they used the program. Teachers in School A averaged about 10 percent less classroom time on population genetics when they used the program. On the average, teachers assigned slightly fewer textbook pages to program classes than to no-program classes.

Table 1

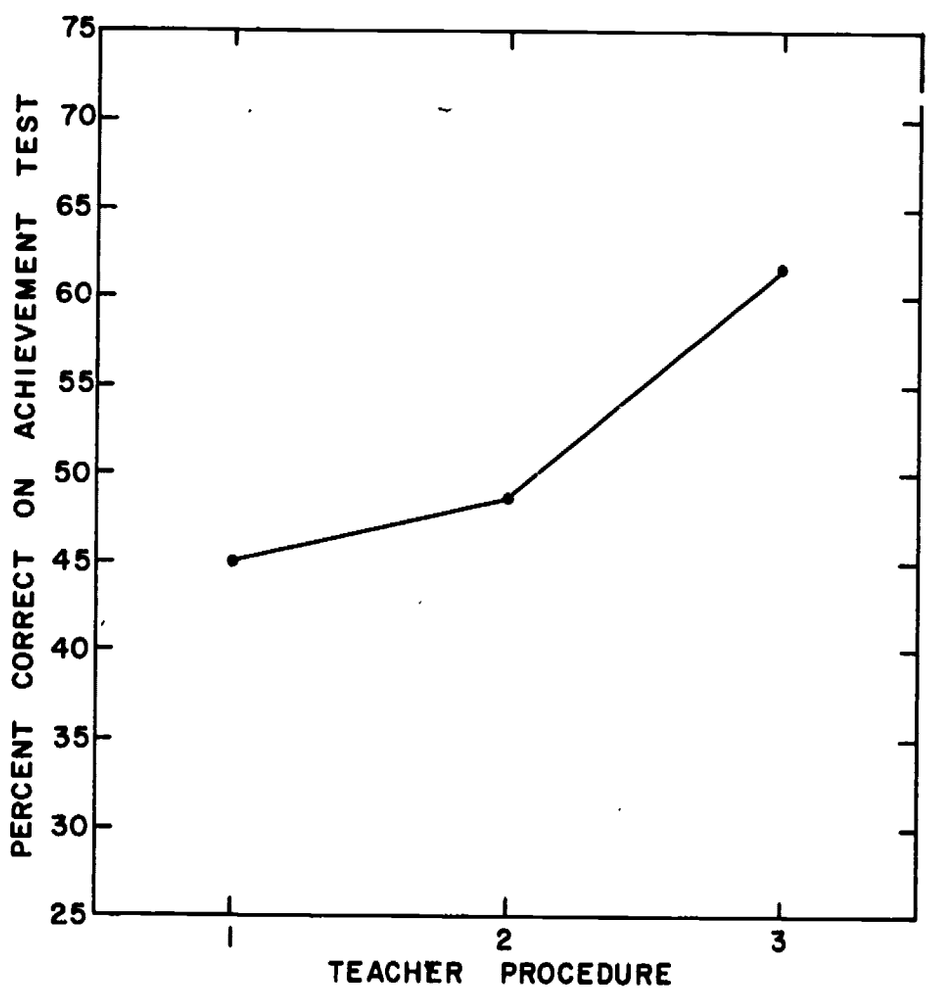
Average Class Time in Minutes by Topic Covered School and Treatment Group

	Program		*No Program*	
	School A	*School B*	*School A*	*School B*
Average amount of class time between pretest and posttest	454	1,031	454	1,031
Time on population genetics				
On program	151	0	0	0
Other population genetics work	52	451	228	451
Total	203	451	228	451
Time on material other than population genetics	251	580	226	580

The teachers were classified according to how they assigned the program. The first group of teachers represented in Figure 4 made the program available, but students were not required to complete it, and no class time was allowed to work on it. For the second group, the program was a required activity but, once again, no class time was provided to work on it. The teachers in the third group reported making the program a definite assignment and providing up to three hours of class time to work on it. However, the data indicate that an average of about four hours is required to complete the program. Fewer than 20 percent reported completing the program in three hours

Figure 4

Achievement as a Function of Teacher Procedure for Using the Program



or less. Thus, most students did not complete the program in class, if they completed it at all.

There is reason to be concerned about the range of results obtained with an instructional package as well as the average result. An F test showed that there was significantly less variability in achievement gain among teachers for classes that received the program than among teachers for classes that did not [$F(7,7) = 6.68$, $p < .05$, two-tailed test]. Moreover, as can be seen in Figure 5, *every* teacher achieved more with the program than without it.*

The teachers were asked whether the program influenced the manner in which they taught nonprogram classes. Three teachers (noted with an arrow in Figure 5) answered in the affirmative. Teacher D said "The organization the teacher used & presentation of the problems [in the program] were used when teaching those classes that had *not* received the program." Teacher H said "I can say that the program helped me to present a better class course in population genetics to all of my classes. I used many features of the program and found them an easier approach to an otherwise difficult topic for many students."

Achievement as a Function of the Student

Anyone developing instruction must make assumptions about what students who will receive the instruction will already know. The population genetics program is based upon the assumption that students who use it are skilled at the arithmetic of proportions and capable of squaring a binomial (and factoring a squared binomial).

There are many educators who believe that if self-instructional programs have any value it is to drill slow students on technical vocabulary. One of the original aims of this project was to demonstrate that programs can be used effectively to teach an interrelated set of concepts and principles to the *brightest* students. It was reasonable to suppose that almost all of the bright students for whom the program was prepared would possess the required mathematics skills. However, it was discovered at a late date that upper-track students, who were to have comprised half the sample in the comparative experiment, were unavailable in large numbers at the time when we

*One teacher is not included in this analysis because of an error in administering forms of the achievement test.

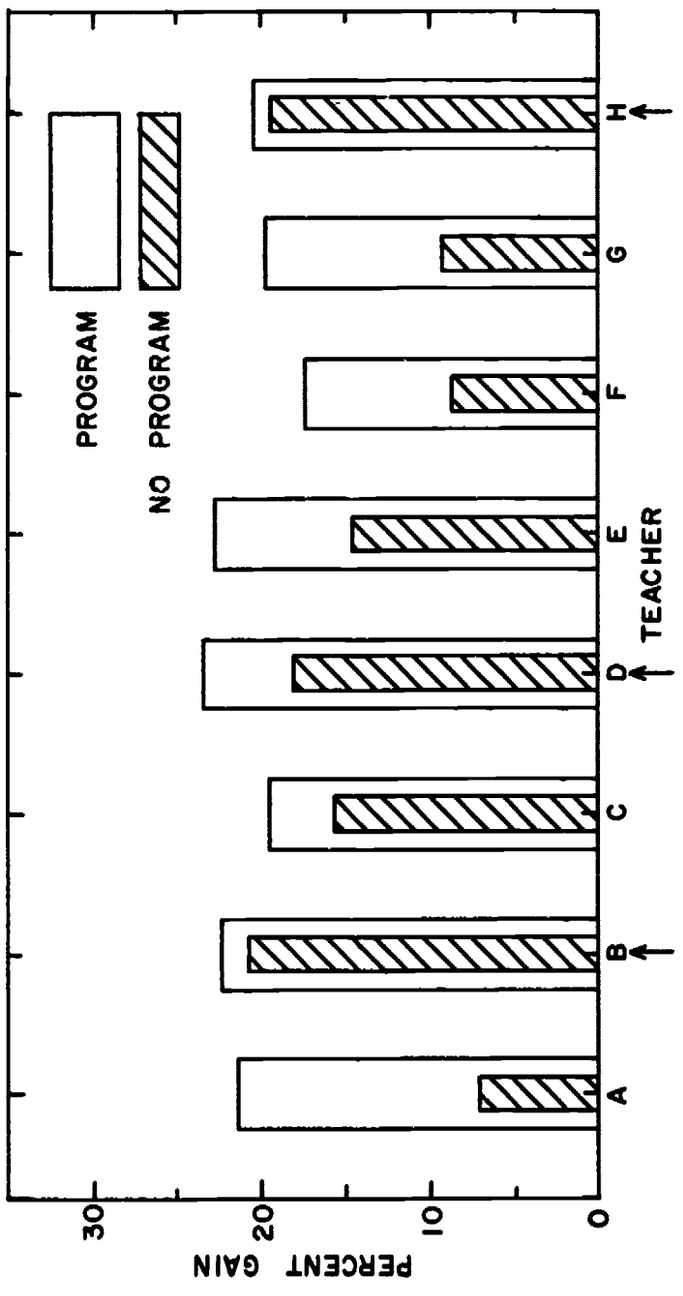


Figure 6
Achievement Gains for Eight Teachers

could be ready to do the research.

A five-item entering behavior test was administered to the students who did participate in the experiment. To our dismay we discovered that only about 40 percent of the sample possessed the requisite mathematics skills—that is, only 40 percent correctly answered at least four of the items on the test. Figure 6 shows posttest performance as a function of entering behavior. Notice that the program was slightly more effective at every level, but its advantage was substantial only for those students who performed well on the entering behavior test. There was a ceiling on the performance of students who did not receive the program, even for those who knew the mathematics.

The questionnaire asked students to indicate how much of the program they actually completed. About 75 percent said they finished the entire program. Naturally, the more of the program that was completed the higher was the achievement. This relationship is diagrammed in Figure 7.

One of the things we do know is that students will not learn much from programs when they simply copy right answers into blanks (10, 4, 7, 11). Students were asked how frequently they turned the page and copied the correct answer when they found a question difficult. Despite the fact that the directions to the program exhorted students to "*write* an answer to each question and fill in each blank *before* you look ahead to the correct answer," better than 40 percent said they sometimes copied answers to difficult questions and 20 percent said they usually did. Figure 8 plots achievement as a function of the frequency with which the student reported peeking at the right answer.

Teacher and Student Attitude

All of the teachers felt the program was a valuable supplement to existing BSCS materials on population genetics. When asked whether they would use the program again, five of the nine teachers indicated they definitely would, two said they probably would, one said he probably would not, and one failed to answer the question. The teachers were enthusiastic about the content and the organization of the program and satisfied with the level of interest it sustained in students. Two teachers volunteered the information that its reputation was so favorable that some students in nonprogram classes borrowed copies of the program from their friends.

Figure 6

Achievement as a Function of Entering Behavior

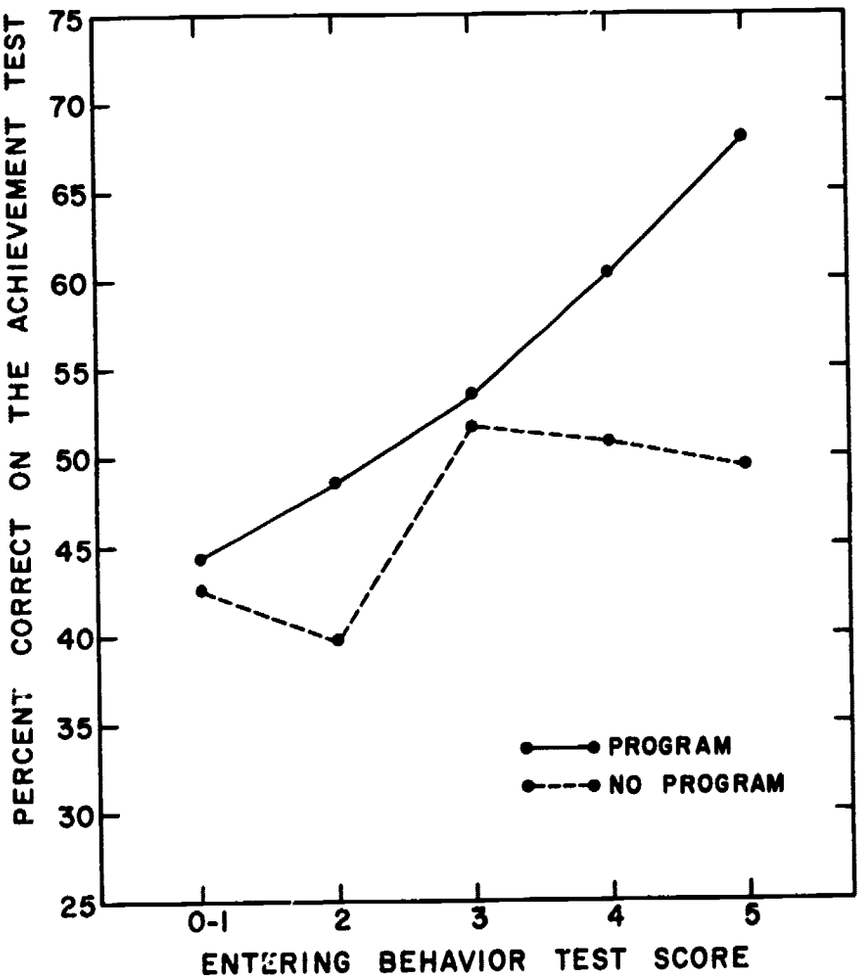


Figure 7

Achievement as a Function of the Amount of the Program Completed

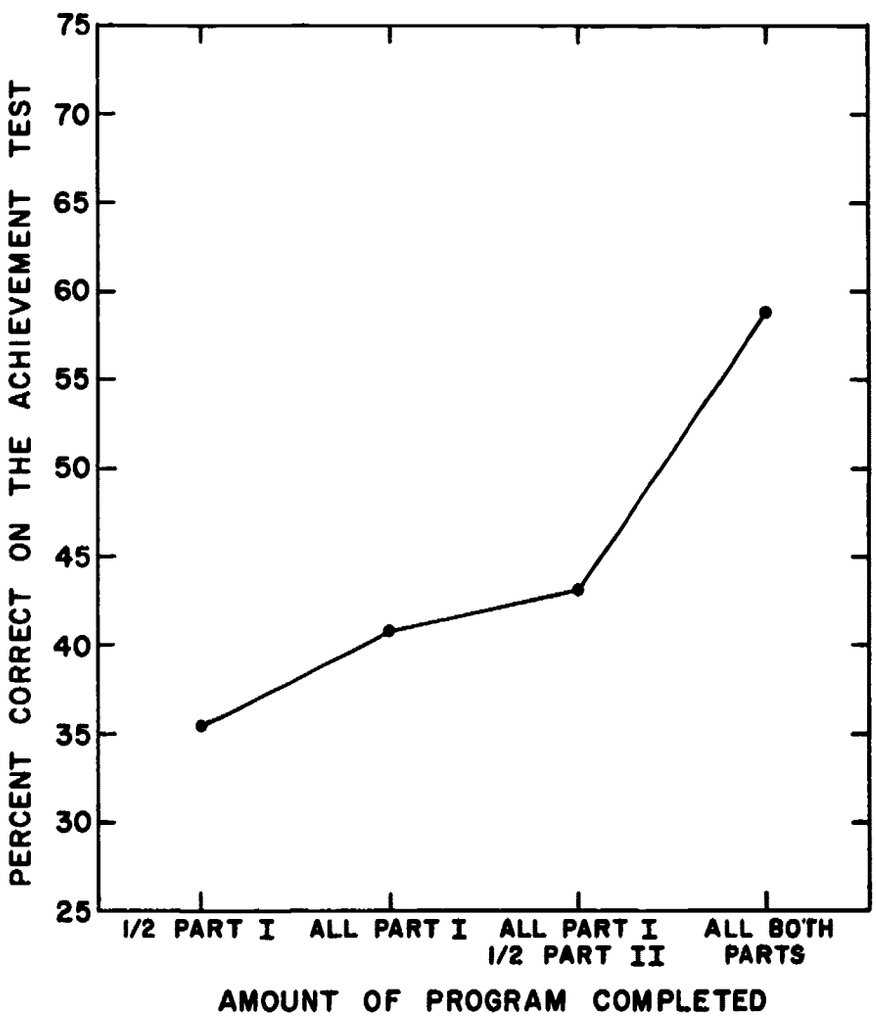


Figure 8

Achievement as a Function of the Reported Frequency of Copying Correct Responses to Difficult Frames

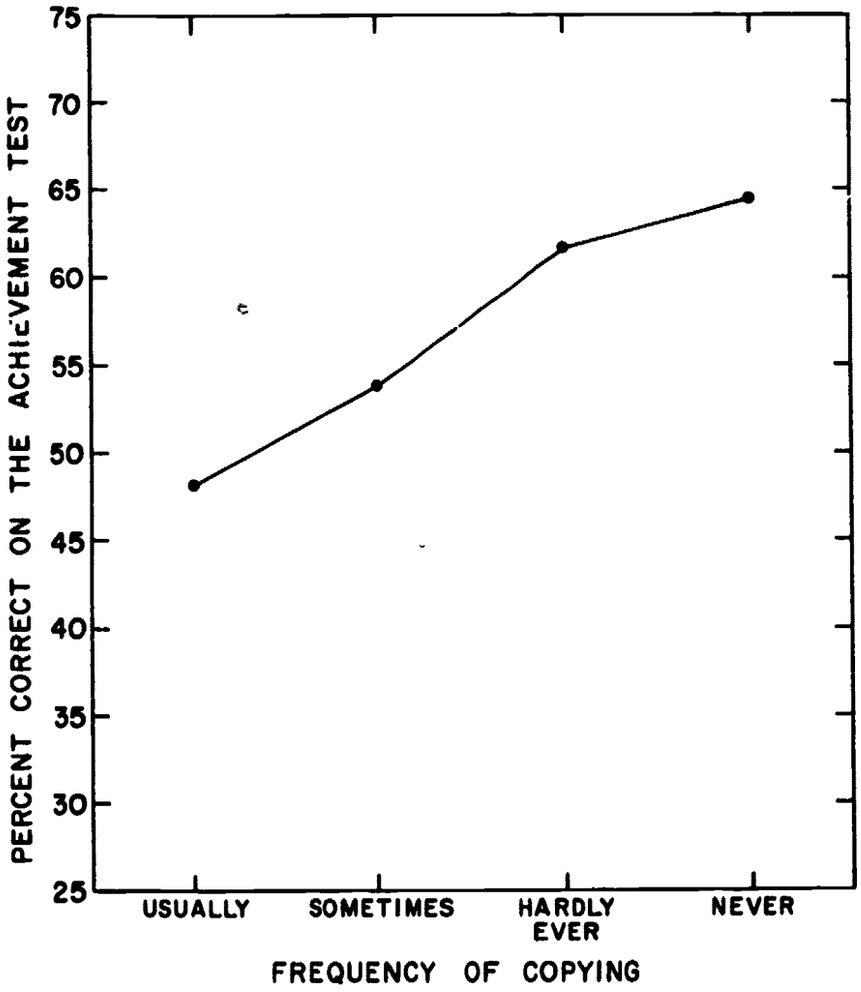


Table 2 summarizes responses to four of the items in the student questionnaire. The majority of the students reported that they would choose to use a program like the population genetics program again if given the chance, that they learn more from this kind of program than from a textbook, and that the program held their interest and required careful thinking.

DISCUSSION

One of the purposes of this paper has been to document the validation of an instructional package. The strategy was to show that the new package, which in this case included a self-instructional program in population genetics, is more effective than a widely accepted and highly regarded comparison treatment.

Instructional effectiveness should be viewed in absolute as well as relative terms. The overall achievement test average for students who (nominally) received the program was 53.6 percent—hardly a satisfactory level. (The average was 43.5 percent for students who did not get the program.)

Under optimum conditions, however, the program led to higher achievement. Students who passed the entering behavior test, who reported that they had completed the program, and who said they never, or hardly ever, peeked at the correct answer achieved an average score of 70.5 percent. The presumption is that the overall achievement average would be higher than found in this study if enough time in class were provided for all students to complete the program; if the program were required instead of optional; if teachers backed up the requirement with the sanctions and reinforcers at their disposal; if students were persuaded to compose an answer to each question before looking ahead at the correct answer; and, of course, if the program were used only with students who possessed the prerequisite skills in mathematics.

Admittedly, an achievement level of 70 percent under optimum conditions is nothing to brag about.* However, in evaluating the level

*Not mentioned previously were three classes of honors students (for whom the program was actually designed) using the BSCS Blue version. The two classes that got the program averaged 83.5 percent on the posttest while the one which did not scored 72.9 percent.

Table 2

Student Evaluation of the Program
N = 377

Percentage of Students	*Item*
	1. If I had a choice
71.6%	(A) I would like to use programs similar to the population genetics program more often.
12.2	(B) I would not care what materials were used.
14.9	(C) I would prefer *not* to use programs.
1.3	Omitted
	2. In comparing a program like the population genetics program with a textbook, I feel that *with the same amount of time and effort*
36.3%	(A) I would learn much more with the program.
42.2	(B) I would learn somewhat more with the program.
8.0	(C) there would be no difference.
10.3	(D) I would learn somewhat more with a textbook.
3.2	(E) I would learn much more with a textbook.
	3. How well did the population genetics program hold your interest?
22.0%	(A) I was very interested.
45.9	(B) I was moderately interested.
22.3	(C) I lost interest at times.
8.5	(D) I got very bored.
1.3	Omitted
	4. To what extent did the population genetics program require careful thinking?
27.9%	(A) Many pages required careful thinking in order to answer the questions correctly.
63.1	(B) Some pages required careful thinking.
5.3	(C) Little thinking was required.
1.9	(D) The program was ridiculously simple and required almost no thinking.
1.9	Omitted

of performance attained, you must remember that as many as 25 percent of the items in the criterion measure test beyond the objectives of the program and that almost 20 percent of the items cover a topic (basic genetics) which is reviewed but not really taught in the program. When considered in this perspective, the achievement produced by the program was not bad whether viewed relative to the comparison treatment or in absolute terms.

The program is currently being revised on the basis of the results obtained in the field test and the criticisms of geneticists and biology teachers.

The general purpose of this paper has been to make a case for the comparative field experiment. A proper humility in consideration of the current state of educational and behavioral science knowledge, a decent regard for the complexities of human learning and instruction, a fair appraisal of the capacity of basic research to increase our power to predict effective instructional configurations together form an argument for a practical strategy of instructional development. The last step in the development process should be a field test to demonstrate empirically the effectiveness of the new instructional package. There is no other way to warrant effectiveness. The chief reason for a field test is to provide data upon which consumers can base a decision to accept or reject the materials. When two curricula share the same goals (or "cover" the same topics), the field test should consist of a comparative experiment. Demonstrating that a new curriculum meets someone's absolute standard of effectiveness is not enough because competing curricula may exceed this standard, may reach the same standard in less time or at lower cost, or may be more acceptable to students and teachers.

Many persons have argued that instruction should be empirically validated. At this point in time there is little indication that anyone has been listening, or having listened, that they have been persuaded. My final word is directed to writers, programmers, curriculum developers, editors, and publishers: You say you have produced a better lesson. Why should anyone believe you? Where is the proof for your claims? Education will take a giant step forward when the producers of instruction empirically validate their products as a matter of standard operating procedure, and when consumers routinely require such validation as a condition for acceptance.

REFERENCES

1. Anderson, R. C. Discussion of instructional variables and learning outcomes. In M. Wittrock and D. Wiley (Eds.), *Problems in the evaluation of instruction.* New York: Holt, Rinehart, & Winston. (In press)
2. Anderson, R. C., and Faust, G. W. *Educational psychology.* New York: Dodd-Mead. (In press)
3. Anderson, R. C., and Faust, G. W. The effects of strong formal prompts in programed instruction. *American Educational Research Journal*, 1967, *4*, 345-352.
4. Anderson, R. C., Faust, G. W., and Roderick, M. Overprompting in programed instruction. *Journal of Educational Psychology*, 1968, *59*, 88-93.
5. Biological Sciences Curriculum Study. *Biological science: an inquiry into life.* New York: Harcourt, Brace, & World, 1963.
6. Bloom, B. S. *Taxonomy of educational objectives.* New York: Longmans, Green, 1956.
7. Brown, R. W. Format location of programed instruction confirmations. *Journal of Programed Instruction*, 1966, *3*, 1-4.
8. Conant, J. B. *Modern science and modern man.* New York: Columbia University Press, 1952.
9. Cronbach, L. Evaluation for course improvement. *Teachers College Record*, 1963, *64*, 672-683.
10. Faust, G. W., Anderson, R. C., Guthrie, J. T., and Drantz, V. E. *Population genetics: A self-instructional program.* I. *Basic concepts.* II. *Genetic stability and change.* Urbana: Training Research Laboratory, University of Illinois, 1967. (mimeo)
11. Kemp, F. D., and Holland, J. G. Blackout ratio and overt responses in programed instruction: resolution of disparate results. *Journal of Educational Psychology*, 1966, *57*, 109-114.
12. Lumsdaine, A. A. Assessing the effectiveness of instructional programs. In R. Glaser (Ed.), *Teaching machines and programed learning, II.* Washington, D. C.: National Education Association, 1965.
13. Scriven, M. The methodology of evaluation. In R. Stake (Ed.), *Perspectives of curriculum evaluation.* Chicago: Rand McNally, 1967. P. 52.
14. Stake, R. E. The countenance of educational evaluation. *Teachers College Record*, 1967, *68*, 523-540.

Evaluation of Teacher Training in a Title III Center

ETHNA R. REID
Exemplary Center for Reading Instruction
Granite School District
Salt Lake City, Utah

The Exemplary Center for Reading Instruction (ECRI) is striving to improve reading instruction in grades K through 12 through demonstrations of exemplary diagnostic and instructional reading programs; through in-service teacher training for developmental, remedial, and clinical reading teachers; by collecting, cataloguing, and disseminating information on materials, training methods, and research; and by maintaining liaison with regional and national research and development projects and related institutions to establish cooperative ventures in program development and research.

In this paper it is impossible to describe all of these activities in detail. Because of the emphasis on in-service training and dissemination services at the Reading Center, I will discuss the following: first, evaluation of beginning reading programs, including materials selection, materials analysis, and teaching-behavior analysis in the use of the materials; second, evaluation of teaching behavior as it relates to classroom management; third, evaluation of the Reading Center's dissemination services; and fourth, basic research as a means of evaluating principles underlying instructional strategy in in-service programs.

STRATEGY FOR BEGINNING READING PROGRAMS

An evaluation strategy aimed at beginning reading programs was developed under the direction of Dr. Gabriel Della-Piana, Director of the Bureau of Educational Research at the University of Utah and University Coordinator at ECRI.

Selecting Materials

Large school districts regularly face the practical task of selecting text materials for beginning reading programs, installing and monitoring them, and revising or adapting these materials or their manner of use to get maximum effectiveness and efficiency. Typically, a state committee selects a list of text materials from which local districts may choose one or more or a combination for trial and adoption in their own schools. After materials are selected, however, district personnel must find ways of installing the materials so they are most effectively and efficiently used. The personnel shortage for supervisory and training tasks of this sort makes it mandatory that materials installation be simplified as much as possible.

To provide a data base for selecting materials, we conducted a local comparative study of beginning reading programs, following this three-step procedure:

1. Review current literature to find out which programs are likely to be maximally effective for specified goals within a district population and select one or two of the most likely prospects. Determine also which programs operating within a district are most widely used there.
2. Conduct a comparative evaluation study of these programs so that the program selected for further evaluation and development will be that which yields the greatest achievement gains on the largest number of outcome measures for all ability levels.
3. Select the program that rates highest in a comparative evaluation and modify it to maximize its effectiveness and efficiency.

The materials-selection phase of the evaluation makes use of a treatments (McGraw-Hill *Programmed Reading* plus other experimental programs and controls) by levels (three levels of beginning-of-year reading readiness) analysis of variance design. The results of this analysis tell us which programs are yielding the greatest end-of-year achievement for different beginning-of-year readiness levels. For example, the *Programmed Reading* treatment yielded greater achievement than did the controls for pupils in the initially high and middle levels on the *Murphy-Durrell Reading Readiness Analysis* but not for pupils in the low level. No single reading program was found to be either significantly better than all others on all variables or to be

uniquely effective for students of any given level of preinstructional readiness. Yet McGraw-Hill *Programmed Reading* was favored most frequently, primarily for the pupils in the high and middle reading-readiness levels.

Some tests, such as the individually administered *Gilmore Oral Reading Test*, could not be administered to all pupils in all programs because of efficiency considerations. Instead of a treatments-by-levels design we made a comparison of total *Programmed Reading* versus total controls on regressed gain scores. Here we found that the *Programmed Reading* pupils were superior to the controls on oral reading rate, comprehension, and accuracy measures, but the oral reading accuracy differences were not significant.

The major conclusion derived from this phase of evaluation was that *Programmed Reading* was generally better than the programs used by the controls (eight basal reading programs), the basal reading programs reinforced with the Educational Developmental Laboratory's machines, and the EDL *Listen Look Learn* system groups. Since *Programmed Reading* was generally superior, attention was focused on improving its effectiveness. Also, since it was found weakest for the initially low-readiness pupils on most measures and specifically on oral reading accuracy, future evaluation was focused on these relatively weak spots.

Although final decisions on which treatment is most effective should await longitudinal studies, some decisions must be made on the basis of available data.

Analyzing Materials

We have noted that in spite of its general superiority, *Programmed Reading* was generally not favored on the oral reading accuracy and other measures among low-readiness students. The focus of our materials analysis was influenced by this observation. One basic question guided our materials analysis: Under the current conditions of use of the materials by teachers and low-readiness pupils, what word-recognition errors (oral reading inaccuracies) occurred consistently in one book of the series and did not disappear or diminish in later books?

Answering this question involved selecting a small group of first grade children who were low achievers, testing them on all words

introduced in the Primer and at the end of Book 1, and testing for these same words at the end of Books 2 through 10. For example, words introduced in Book 2 would be tested at the end of Book 2 and also at the end of Books 3 through 10. Table 1 gives a sample of one child's responses to words introduced in the Primer and tested at the end of Books 1 through 5.

Table 1 shows that some words are never missed (a, I, man, pan, yes); some are missed each time they are given from Book 1 through Book 5 (ant); some are missed one or more times and then no more (fat, mat, pin), and others are correct at first and then later missed (no, thin, am). Collecting these data from a group of pupils provides the direction for developing supplementary instructional material and teacher programs. For example, consider the word "fast" introduced in Book 1 and tested at the end of Books 1 through 5. A summary of the errors is presented in Table 2.

The most common errors on words introduced in Book 1 and persisting through Book 5 are: that/did, dig/did, fins/fans, fins/fit, fat/fit, him/ham, hat/hit, ing/in, mitt/mint, Mrs./Miss, pants/pant, pat/pant, pant/pats, pants/pats, sand/sad, sad/sat, sit/sat, sting/sing, sat/sit, thin/this.

The following recommendations are based on the type of data presented in Tables 1 and 2:

1. It is probably inefficient to contemplate any program modification or material supplements for words that are missed frequently in Books 1 and 2 and thereafter never missed.

2. It would be wise to analyze the determinants of word errors which persist over a span of four or five books. For example, the most frequent erroneous readings of "fast" were "fat," "fats," or "fit." An analysis of the test material in Book 1 shows that several items requiring discrimination between "fat" and "fast" are presented (for example, a picture of a thin fish and the statement "this fish is fat/fast"). Since the discrimination problem persists, perhaps some supplementary material should be designed without picture clues for pupils missing this item at the end of Book 1. Alternatively, perhaps the pictures should be modified to prompt the correct response.

The reason *Programmed Reading* was not statistically superior in a significant way to other treatments on oral reading accuracy is prob-

Table 1

Record of one Student's Oral Reading Errors on the PROGRAMMED READING *Word List Across Five Testing Periods*

Student's Name											
Date		2-7	2-21	3-13	3-29	4-24					
Book		1	2	3	4	5	6	7	8	9	10
End-of-Book Test	Word										
1	a										
1	am(2)*		the the			man					
1	on		am	ont		am					
1	ont(2)	ot hot	pont hands	hot and	and and	and					
2**	ant			and	hont	pont					
1	I										
4	I										
1	fat	has	fit								
6	fat										
1	man										
2	man										
5	man										
1	mat		hap								
1	no		on	on	on						
3	no										
1	pan (3)										
2	pan										
5	pan										
1	pin		pins								
5	pins										
1	tan					tin					
1	thin		him	the	that	this					
2	thin		D.K	that	that						
4	thin				that						
5	thin										
7	thin										
1	yes										

*Number in parentheses refers to the number of times the word appears in the test.
**"Ant" appears in end-of-book test 1 and again in end-of-book test 2. Similar circumstances hold wherever a word is followed by blackened squares within the table (for example, the word "fat" does not appear as a test word a second time until end-of-book test 6).

Table 2

*Error Record for the Word "Fast," Introduced in Book 1 and Tested for at the End of Books 1 through 5**

Book	1	2	3	4	5
Number of Pupils	7	10	12	11	11
Errors	fat (2) fin it (2) fit dish	sang (2) sat	went fats (2) D.K. past.	pig fit	fats fin

*McGraw-Hill PROGRAMMED READING

ably because of the discrimination problems caused by the high dependence of *Programmed Reading* materials on picture cues and grammatical sequence as prompts for filling in blanks. These limitations in materials design can be solved by providing supplementary material or by changing the teacher's use of the regular text materials.

Analysis of Teaching Behavior

At the Exemplary Center for Reading Instruction, we are taking three approaches to the analysis of teaching behavior to reveal deficiencies that may account for the lack of superiority of *Programmed Reading* over controls for the low-readiness pupils. These three approaches are: 1. identifying the most and least effective teachers for the low-ability pupils; 2. observing the behavior of these teachers in using the text materials; and 3. developing observational systems for detecting effective pupil-management techniques.

Identifying differences in teacher effectiveness: By the time we began to identify teacher effectiveness, we discarded the control group be-

cause the *Programmed Reading* classes were not poorer than the controls on any variable. Our objective was to maximize the effectiveness of one of the "best" reading programs in use. So we limited our analysis at this stage to distinguishing between teachers in the *Programmed Reading* group with respect to gains of low-ability pupils. The procedure was as follows:

1. Compute a regression equation for the September (Murphy-Durrell) to May (Gilmore Oral Form B) data for all 89 pupils in the low Murphy-Durrell group in *Programmed Reading*.
2. Determine the residual gain score for each pupil (difference between his predicted May score and his actual May score).
3. For each teacher, tally the number of low Murphy-Durrell pupils who fall above the regression line (perform better than predicted), the number who fall below the line, and the number (if any) who fall on the line.
4. Those teachers who have the greater percentage of their low-ability pupils above the line are the ones who are producing the greatest gains for that ability group.

The data for the 12 *Programmed Reading* teachers in our sample are presented in Table 3. How may these data be used? First of all, for confident decisions about such teacher differences we would want to measure and observe over a period of two or three years. Nevertheless, since we must make some decisions concerning program improvement each year, the data can be put to an immediate use. Teachers 4, 6, and 8 have more students achieving scores below the regression line. The regression line is based on the correlation between beginning-of-year readiness scores and May reading achievement for all 12 classes combined. Teachers 7 and 9 have more students above the regression line. Yet if we look more carefully at other data on these classes, we find that teacher 9 has only 2 low-ability pupils as compared with 4 to 11 for other teachers The teaching methods used by teachers 7 and 8 could be compared, since they taught equal numbers of low-ability pupils whose socioeconomic characteristics were judged to be identical (barring bias in forming classes) by virtue of their attending the same school. This approach identifies teachers producing the greatest percentage of low-ability pupils who score above the regression line. Once these teachers are identified, they can

be observed to determine which characteristics of their behavior are responsible for the differences in effectiveness.

Comparative data on the instructional methods used by teachers 7 and 8 would be useful to guide program improvement. However, it would be more desirable to have data on a larger number of teachers. The ideal procedure would be to begin with a larger number of classes in schools each having three or four first grade classes and where there was random assignment of children to classes.

Table 3

Number of Low Reading Readiness Students Falling Above, On, or Below the Regression Line for Predicting May Vocabulary and Comprehension Scores on the* GATES-MACGINITIE READING TEST

			Above		*On*		*Below*	
Teacher	School	Group Size	Vocabulary	Comprehension	Vocabulary	Comprehension	Vocabulary	Comprehension
1	A	10	4	5	—	—	6	5
2	A	9	4	5	—	—	5	4
3	B	7	4	4	—	—	3	3
4	B	11	3	4	1	0	7	7
5	B	7	4	2	—	—	3	5
6	B	8	2	2	—	—	6	6
7	C	6	6	5	—	—	0	1
8	C	6	2	1	—	—	4	5
9	D	2	2	2	—	—	0	0
10	E	9	3	4	—	—	6	5
11	E	4	2	1	—	—	2	3
12	E	10	4	4	—	—	6	6

*Scores ranging from 16-73 on the MURPHY-DURRELL READING READINESS ANALYSIS, September 1968

Observing teaching behavior: The degree to which all teachers used the teacher's guide was observed during the materials-selection phase of evaluation. Teacher 7 (who obtained better-than-average gains for low-ability pupils) deviated more from the teacher's guide and spent more time with low-ability pupils than did teacher 8. Teacher 8 was rated "1" (high fidelity to the guide) and teacher 7 was rated "4" (low fidelity to the guide) on a five-point scale. Both of these teachers moved at the end of the first year's study, so we were not able to observe their behavior on other dimensions. Table 4 lists some of the dimensions of teaching behavior which we see as relevant.

Note that we have included a category for observation of the pupils' behavior, since an adequate description of how a teacher manages the class must include some of these data. More detailed behavioral data on pupils can be obtained by procedures which will be discussed later.

Informal observations made while using the teaching-behavior checklist have identified some variability in teaching behavior including a rather common lack of detailed diagnostic and prescriptive teaching, a defect which should significantly affect the performance of slower pupils. At any rate, behavioral data can be applied toward the goal of maximizing a reading program's effectiveness by identifying behavior characteristics which discriminate between high- and low-gain teachers of low-ability pupils, by obtaining base rates on significant teacher and pupil behavior, and by providing direction for in-service training programs. The base rates, of course, may be used as a baseline against which teaching behavior observed during and following teacher training can be compared.

Teaching-observation systems currently being developed: An example of an observation system for one aspect of teaching behavior listed in Table 4 (*reinforcement contingent on performance*) is presented in appendixes A and B. The system is designed for observing and recording the extent to which teachers establish (promise) contingent stimuli ("If all of you finish your work before 10:30 we will play our spelling game.") and apply them (actually carry out the contingencies promised). The system is broad enough to deal with positive reinforcers, punishment and escape contingencies, and application of contingencies not previously promised.

The system's sensitivity to teaching differences was demonstrated in a pilot study. Five teachers trained in contingency management were compared with 14 untrained teachers observed in a previous study. The five teachers were observed in four separate half-hour

Table 4

Categories for Observation Schedule for Teacher and Pupil Behavior (Low-ability Pupils) in Teaching with McGraw-Hill PROGRAMMED READING *Materials**

Teacher Behavior

—end-of-book test
- administers
- listens to oral reading
- records errors
  - accuracy of recording
- records causes of errors
  - accuracy of causes

—teacher time on task
- percent of reading time teacher not diagnosing, prescribing, or teaching

—making prescriptions of objectives (written or mental)
- related to diagnostic data
- specific response described
- situation in which response is specified
- criteria for an acceptable response specified

—prescriptive teaching (described/conducted)
- achievable bits
- prompts for evocation of response
- feedback
- fade prompts
- overlearning
- varied context practice
- related to diagnosis
- reinforcement contingent on performance

—fidelity to teacher's guide (to be listed in detail)

Pupil Behavior

—number of pupils in class/number with *Murphy-Durrell* score 73 or below

—each child: workbook number/date

—each child: total time allocated to reading period/percent time at task

—each child: time allocated to independent reading/percent time at task

*Instructions for time sampling of behavior not included

sessions at the beginning of the reading period. The results indicated that:

1. Verbal positive contingent reinforcement was more frequent for trained teachers (28 per half-hour) than for untrained teachers (11 per half-hour).
2. Verbal negative contingent reinforcement was more frequent for trained teachers (12 per half-hour) than for untrained teachers (3 per half-hour).

Thus, teachers may be observed for the categories indicated, differences between high and low gain-producing teachers may be observed, and training programs may be initiated to produce desired changes in teaching behavior.

EVALUATION STRATEGY FOR CLASSROOM MANAGEMENT BEHAVIOR

The current emphasis within the realm of individualized instruction is upon the relationship between student work skills and teaching behavior. Individualized instruction requires that the teacher spend time with each child on his program or his products. The teacher's role becomes that of expert in diagnosis, prescription preparation, and general trouble shooting. Ideally, with no pragmatic demands, this might be met by a one-to-one tutorial setting, at least for much, if not all, academic instruction. Realistically, it must be met by the teacher moving from child to child as needed, yet maintaining a productive classroom. That is, children should be able to work individually, yet get help when their progress is thwarted.

In many classrooms the teacher decides when his pupils should work; he continually prods them to attend, to continue working, and to complete what they are doing. He keeps them moving from subject to subject and from unit to unit. Unfortunately, these procedures are not always effective in maintaining a productive classroom.

The reasons for inadequate development of self-controlled work skills are probably numerous. Ideally, we would like children to continue working without constant teacher intervention. Yet many

teachers quickly interrupt when a student stops working, and ignore him when he is working satisfactorily. These teachers act as if their task is to detect and correct "misbehavior" only. Such differential attention might inadvertently reinforce idleness or misbehavior. For example, Wesley C. Becker, University of Illinois (unpublished), experimentally demonstrated in a classroom that the more the teacher told children who were out of their seats to sit down, the more they left their seats. Teacher reinforcement of other poor work skills is probably also common.

Another problem concerns the kind of reinforcement contingencies teachers use to maintain work skills. In some classes the essential "control" procedure is an "escape" contingency. That is, rather than positively reinforcing appropriate behavior, the teacher bombards his students with repeated instructions, threats, and criticisms when they are not working. The children go to work or do whatever is necessary to terminate, and thus escape, their teacher's unpleasantness. The verbal barrage ceases when the children return to work. Even though it gets students to work, such a procedure is likely to fail in the long run because it develops no motivation to work other than the motivation to escape the noise. When the threats and directions stop, the work may also stop. The teacher is then trapped in a predicament in which he must continually repeat the instructions, threats, and demands if he is to maintain student behavior.

One phase of ECRI's in-service teacher-training programs includes the development of innovative practices in strategically located "Skill and Product Development Classrooms" which serve as exemplary supply depots from which area training and demonstration teachers set up similar programs in other schools.

Practical, easily taught techniques for teachers in establishing and maintaining students' independent work skills constitute the primary objective of one of the SPD classrooms. The observation system for evaluating teaching behavior in the classroom has been developed and is currently being refined under the direction of Dr. Howard N. Sloane, Jr., University of Utah.

Program Preparation and Reinforcement System

Two classes of independent work-skills behavior have been identified. The first concerns paying attention to instructions and not engaging in

competing behaviors while the teacher is giving instructions. The second relates to pupil maintenance of independent work. Specific programs to develop each of these classes of behavior are being used.

A reinforcement system requiring very little teacher effort has been designed. Students can earn points for the class as a whole during classroom activities in which the teacher must interact with several students simultaneously. The teacher tallies these points on an inexpensive, electronic counter which students can see and hear from anywhere in the classroom. Students can earn individual points for correctly completing an academic unit, and they may periodically trade their points for "back-up reinforcers" such as classroom privileges and activities.

Evaluation Design

The major evaluation instrument for student and teacher behavior has been developed, and its reliability is now being checked. Appendix C includes procedures for coding major areas of student and teacher behavior and a summary copy of the classroom behavior scale.

Through the use of this instrument by trained observers, teacher and student behavior changes can be evaluated by observation and rating before and after training sessions. Initial reliability data on the classroom behavior scale are included in Table 5.

Table 5 shows the percent of agreement on the classroom behavior scale among three observers who had been trained over a three-month period. The observers rated five randomly selected second grade students over nine thirty-second observation periods as described in the rating procedures.

Percent agreement was calculated as:

$$\text{Percent agreement} = \frac{\text{number of agreements}}{\text{number of agreements} + \text{number of disagreements}}$$

In assessing agreement and disagreement, different rating codes (0's and I's for example) applied to the same behavior by different raters within a single rating interval constituted disagreement.

Raters indicated that they recognized a behavior's absence by drawing a diagonal line across all rating intervals. The use of a diagonal line left no question as to whether raters were sure the behavior

Table 5

Estimate of Interobserver Agreement on the Classroom Behavior Scale

Behavioral Subclasses		*Percent of Agreement*
AREA 1	SB	
	U/D	87
AREA 2	NV-T	
	p	100
	c	100
	h	99
	a	100
	+	99
	−	99
AREA 3	V-T	
	+	100
	−	99
	I	99
	O	100
AREA 4	V-O	
	+	91
	−	91
	I	69
	O	88
AREA 5	TI	
	i	75
	g	94
	c	66
	◯	63
	'	89
	−	77

in question did or did not occur. Therefore, agreement in using the diagonal line constituted agreement in using the rating codes which it replaced on the scoring records.

Additional evaluation data will consist of records of assignment completion and correctness, measures of student attending behavior, and some general measures of academic achievement. The demonstration class and experimental (or field training) classes will be compared with themselves and with other control classes.

DISSEMINATION EVALUATION

The Exemplary Center for Reading Instruction maintains four distinct dissemination avenues, each requiring independent evaluation. The four include: (a) out-of-Center services such as those rendered by the area training teachers; (b) library loans; (c) demonstrative and exemplary functions within the Center, including visits and tours; and (d) dissemination via the mail service. Evaluation programs for (c) and (d) are in progress under the direction of Dr. Jon E. Atzet, Reading Center psychologist and co-editor of the ECRI *Newsletter*.

Dissemination via the Area Training Teacher

This dissemination medium involves the Skill and Product Development Classrooms described earlier and ECRI's area demonstration training teachers. The area training teachers instruct classroom teachers in individual prescription techniques, in establishing effective independent work skills among their students, and in teaching their students an elemental approach to critical reading. The teachers are the targets; it is their teaching techniques and behavior management that are to be modified and honed to raise classroom efficiency and productivity.

Dissemination via Library Loan

The library faces a unique problem in evaluating its program. Though

its holdings make up several distinct categories or subject areas, its patrons are individuals who frequent the library to fill their particular needs. Several patrons might use and assess the worth of a certain item, but the premium each places upon it will differ from person to person and will be determined by the unique way in which each person uses it. Library holdings, therefore, have no indigenous function and cannot be compared against a success criterion. Service, on the other hand, can be.

Circulation records reflect demand, and demand reflects worth or value. Since functionally similar holdings are categorized together in the library's organization, each category's intrinsic value can be determined by statistical comparisons of circulation records among categories. Further, circulation records for all categories can be combined to establish periodic service records for the library as a whole.

Librarians do not normally tally the number of inquiries they receive about materials their facility does not have available. Yet such inquiries demonstrate an interest in certain materials and can, therefore, be used as evaluative data. At ECRI such inquiries are recorded, categorized, and tallied so that they reflect interest in materials not available through the library, and materials which are frequently requested are later installed in the library.

The library evaluation program has also been directed toward estimating the demand for its services. Requests reach the Center by mail, by telephone, and in person. All request letters are filed; telephone calls come through the front desk and are rerouted to their ultimate destination where their messages are recorded and filed until they are counted and categorized. Requests filled in person are estimated from depleted materials stores and from library circulation records.

During the first quarter (January-March 1968), readers requested 107 copies of articles reviewed in the ECRI *Newsletter*, over 300 *Library Resources* books were sold, visitors took away tens of thousands of the many teacher aids and in-service training pamphlets and bulletins, and library circulation reached 33,713.

Dissemination via Visits to ECRI

Reading Center visitors participate in demonstrations, workshops, and lectures; use the library; consult with teachers and other per-

sonnel; tour the Center; and participate in ECRI-sponsored functions held outside the Center.

Data from the questionnaire that we circulate among visitors reveal that during the first quarter (January-March 1968), approximately 59 percent of the Reading Center visitors were teachers, 15 percent were educational administrators, and 26 percent were from other occupations ranging from university students to commercial welders. Seventy-five percent of the visitors came from within the Rocky Mountain region and 25 percent came from outlying states.

The reasons which Reading Center visitors listed for their visits indicate that 6.5 percent came because their children were being instructed in the Reading Clinic, 42.5 percent came to participate in either demonstrations or workshops, 5.9 percent came to use or learn about the Reading Center library, 48.7 percent came to tour the Reading Center or for a general introduction to its facilities and functions, and 2.5 percent came for unstated reasons. Several people came for a variety of reasons and were therefore included in more than one tally.

ECRI's influence extends far beyond the Rocky Mountain region. It has served all of the continental United States, Alaska, Hawaii, and parts of Canada. Workshop participants, consultants, and visitors to ECRI have represented 29 states including Hawaii and Alaska.

Dissemination via the Mail Service

Through the mail service, ECRI dispenses such items as newsletters, bulletins, professional reports, and announcements. Evaluating this medium can be difficult because there is no personal contact between the disseminating and the target agencies. If readers are to evaluate the mailings they have received, a follow-up effort must be made to reach them.

A follow-up evaluation program is expensive. Besides the cost of two-way postage, a self-explanatory, mail-sized evaluation form would have to be designed, mailed, returned, and sorted, and its contents tallied, all of which would consume many costly man-hours.

A follow-up evaluation program makes additional demands upon the evaluator. He is asked to evaluate materials he read some time ago. Provided the reading material has not been misplaced or discarded, the evaluator must refresh his memory on pertinent points by

rereading and pondering the information in the light of the questions on the evaluation form. Then he must complete the evaluation form and return it.

Many potential evaluators habitually shun evaluation programs because of earlier experiences with demanding questionnaires. Others faithfully comply by filling out an evaluation form but neglect the more important task of preparing themselves to do so. Thus they sabotage evaluation accuracy and utility.

Follow-up evaluation programs are often weak because of insufficient compliance. In instances of indirect confrontation such as the follow-up evaluation, compliance is generally inversely proportional to the effort demanded by the evaluation questionnaire unless it is controlled by an attractive form of reward. But in this case the reward carries an intangible, and too often valueless, "do-it-for-science" flavor.

Evaluation relevancy often suffers because potential evaluators are not adequately qualified. Comparison is a fundamental part of evaluation. Its application is exemplified in the before-after and the experimental-control techniques used in scientific investigations. The need for comparison in evaluation imposes the qualification that an evaluator must be well acquainted with the material—and similar materials—he is to appraise. The more extensive his familiarity with related materials, the better equipped he is as an evaluator.

If comparison in evaluation is infeasible, as might be the case with certain innovations which neither replace nor resemble other methods or materials, more stringent qualifications are demanded of the evaluator. He must be willing to take the time to survey the material carefully, looking for inherent merit, potential alternatives, and potential pitfalls. The evaluation, in this case, must reflect exclusively upon the materials being evaluated.

The ill-equipped evaluator tends to shower his subject with praise, virtue, flattery, and so on, which translate into positive or favorable evaluative data. This "halo effect" is tremendously effective in boosting self-estimations, but such evaluation returns do not reflect upon the actual quality of the program they were intended to assess.

Follow-up evaluation program: An evaluation program should be extensive enough to disclose a program's inadequacies, shortcomings, minor faults, and, of course, its strong points. A program design, free from internal difficulties, will provide more latitude for approaching the evaluation program's purpose.

Most problems encountered in a follow-up evaluation program can be controlled through questionnaire design. First, an evaluation questionnaire should be small enough to be sent in a regular mailing. Mail pieces and evaluation questionnaires of equal size can be sent as a unit, saving half of the postage which must be spent if questionnaire size demands that it be sent separately. Incorporating a questionnaire into a standard mailing minimizes the time lag between reading and evaluating. It alerts the reader-evaluator to read with the evaluation objective in mind. The questionnaire guides his reading and prevents him from having to review in order to appraise materials that he once read. Eliminating the time lag between reading and evaluating increases evaluation validity by reducing memory loss and thus the "halo effect" that is most prevalent under conditions of ignorance and/or failing recall. The greatest benefit, however, is that a reduction in effort can increase compliance.

When a questionnaire accompanies any disseminated material, it becomes feasible to gather evaluative data randomly on each mailing rather than from selected readers at selected times. The selection factor alone can reduce evaluation validity because selection is directly contrary to scientific sampling methods.

Second, the questionnaire should carry only those questions which are most relevant to evaluation. The questions should be concise, terse; none of them should be open-ended. Well-structured questions shorten a questionnaire's length and complexity, thereby expediting both the response to it and the tallying of data from it once it is returned. Structured questions provide the evaluator with an evaluation guide; they prevent him from having to conjure up his own evaluation categories. Structured questions provide the evaluator with a type of reading guide enabling him to read for evaluation as well as for his own purposes.

Questions can be leveled at the evaluator's qualification level or designed to control depth of thought. Where evaluation validity is a major concern, question and questionnaire structure can counterbalance lack of qualification.

Third, an evaluation questionnaire should carry a brief description of its purpose stated in such a way as to emphasize the importance of evaluation and the contribution made by each evaluator. The emphasis should be directed at increasing compliance and conscientiousness of effort.

Evaluating the ECRI *Newsletter:* Of the countless objectives that

could be tied to an educational publication, only those that reflect the publishing organization's expressed purpose and its readers' needs should be considered in establishing the criteria against which the publication is to be evaluated. Even though the readers' needs appear to be the primary concern, they are not so important that they should be allowed to alter the publishing organization's purpose. For example, readers cannot legitimately demand information that lies outside the publisher's domain and complain if they do not get it. From the outset, then, the publishing organization is obliged to state its objectives clearly and publicly, and to serve its readers within the limits established by these objectives. Evaluation is the process by which readers assess, primarily, the degree to which a publication's contents actually reflect its objectives, and, secondarily, the degree to which the publication fills their own needs.

The *Newsletter's* objectives, as formulated by the ECRI staff, were: 1. to disseminate information on the Reading Center's functions; 2. to disseminate reading-research findings derived from studies sponsored by the Reading Center; 3. to provide teachers with effective exemplary practices and classroom aids; 4. to provide educators with a medium for publicly commenting on current practices and innovations in teaching reading; and 5. to help educators keep abreast of changes in teaching techniques, materials, and educational philosophies.

In pursuit of its objectives, each issue of the ECRI *Newsletter* features a progress report on ongoing reading-related research sponsored by the Reading Center (shown in Table 6 as Section A) and carries a synopsis about the author (Section B). A third part of the *Newsletter* (Section C) provides a detailed description of an exemplary teaching practice. Section D is reserved for readers who wish to comment on the *Newsletter*'s content and related issues. Section E reports on ECRI-sponsored projects other than concurrent research. Section F presents a review of pertinent, recent research in reading from throughout the world and offers these reports in their entirety through ECRI upon request. Section G provides short, concise suggestions for increasing motivation to read. Sections H and I refer respectively to the cartoons and photographs which supplement the text.

Designing an evaluation form: An evaluation study was undertaken to determine how effectively the *Newsletter*'s contents are fulfilling its objectives and satisfying its readers. An extensive evaluation questionnaire was developed and repeatedly condensed until it fit on one side of an 8½-inch x 11-inch sheet. The other side of the sheet was divided

horizontally in two. The reasons for the evaluation and the general directions were printed on one half; a self-addressed, postage paid, return cover filled the other. Appendix D contains a sample copy of the evaluation questionnaire. The evaluation form was mailed with an evaluation issue of the ECRI *Newsletter*.

Item I on the questionnaire allowed those on the mailing list to either continue or discontinue their subscriptions by checking the appropriate box and returning the questionnaire. ECRI originally adopted the policy of mailing the *Newsletter* regularly to everyone on its mailing list. This policy was to guarantee all potential readers an opportunity to experience the *Newsletter*'s impact, to develop a personal interest in it, and perhaps to pass a copy on to others who might share their interest and submit requests for the publication. The policy has worked well. More than 350 subscriptions were received from March to June 1968 from readers who were introduced to the *Newsletter* by friends.

Because those who are not educators as well as educational administrators who are far removed from the classroom were also represented in the swelling 7,000-entry mailing list, circulation probably exceeded readership. Their actual interest in the *Newsletter* was probably very low, and because of the disinterest, it was expected that some of them would withdraw their subscriptions.

Item I was included in the evaluation questionnaire also to separate *Newsletter* readers from nonreaders to preclude using nonreaders' data in the evaluation.

Item II was incorporated into the evaluation questionnaire: 1. to measure the relative extent to which each section of the newsletter was read; 2. to determine which sections readers thought should be given more space or emphasized for their benefit; and 3. to isolate those sections which were of little or no value to readers. Such information was to guide the editors in redesigning the *Newsletter*'s format to satisfy its readers' needs more effectively.

Item III represented an effort to generate ordinal data which could be applied toward a minute and exhaustive evaluation of individual *Newsletter* sections. Readers were asked to rate (on a 1-5 scale representing excellent through poor) sections A through I on their relative clarity, informativeness, interest value, importance, utility, applicability, practicality, originality, and influence—a variety of attributes that could be applied indirectly to an assessment of the objectives outlined for the publication as a whole.

The purpose of Item IV was to identify and categorize all occupations represented in the readership. Grouping by readership was to provide the third dimension for the data analysis.

Item V provided evaluators with space to comment freely on the *Newsletter*. Open-ended responses could supply relevant, qualifying information not allowed for elsewhere in the evaluation questionnaire.

Results: The response to Item IV demonstrated that the ECRI mailing list contains a diverse sample of the educational populace, a full cross section of people representing education in one way or another. Several distinct groups emerged from the response sample: elementary school teachers (shown in Tables 9-11 as ET), elementary school administrators (EA), secondary school teachers and administrators (STA), college and university teachers and administrators (CTA), and a fifth group comprised of other administrators and educational specialists (O).

Inasmuch as each of the above groups was thought to have a somewhat different professional mission, it was hypothesized that each would assess the various sections or the entire *Newsletter* differently. Table 6, in presenting the analysis of the response to Item III, shows this hypothesis to be a misconception; GROUPS did not differ significantly among themselves ($p > .10$). Neither were SECTIONS by GROUPS nor RATINGS by GROUPS interactions significant ($p > .25$). These results mean that readership groups did not differ among themselves in the way each of them rated the *Newsletter* sections and used the rating categories.

Collectively, however, groups rated each of the *Newsletter* sections and applied each of the rating categories differently (Table 6; SECTIONS, RATINGS, and SECTIONS by RATINGS; $p < .01$). In Table 7, the *Newsletter* sections are ranked according to the magnitude of the overall rating score received by each. Exemplary Teaching Practices (C) ranked highest; then came Reading Research Review (F), Feature Article (A), Reading Keys (G), ECRI-sponsored Projects (E), Letters to the Editor (D), Photographs (I), Cartoons (H), and About the Author (B). All ordered pairs of rankings, except F and C, E and G, I and D, H and D, and H and I, were significantly different from one another (Table 7).

In Table 8, the rating categories are ranked according to their individual total scores—the sum of all numerical ratings for each rating category contributed by all readership groups across all *News-*

letter sections. The *Newsletter* was rated highest on clarity (12); then came informativeness (11), interest value (10), importance (9), utility potential (8), originality (5), influence (4), practicality (6), and applicability (7), respectively. All ordered pairs of ranking, except 8 and 9, 5 and 9, 5 and 8, 4 and 8, 4 and 5, 6 and 8, 6 and 5, and 6 and 4, were significantly different from one another.

Newsletter sections were further ranked from the data received in Item II according to sections read most regularly (readership strength). The Feature Article was the most heavily read (Table 9). Exemplary Teaching Practices, Reading Research Review, ECRI-sponsored Projects, About the Author, Reading Keys, Letters to the Editor, Photo-

Table 6

Comparison of Preferences for and Ratings of Nine* NEWSLETTER *Content Sections by Five Readership Groups*

Variance source	*df*	*MS*	*F*	*p*
GROUPS	4	35.58	2.25	NS*
ERROR	20	15.78		
SECTIONS	8	35.11	12.36	<.01**
SECTIONS × GROUPS	32	2.64	.92	NS
ERROR	150	2.34		
RATINGS	8	2.62	4.12	<.01**
RATINGS × GROUPS	32	.504	.79	NS
ERROR	160	.636		
SECTIONS × RATINGS	64	.561	2.47	<.01
SECTIONS × RATINGS × GROUPS	256	.224	.98	NS
ERROR	1,280	.227		

*The data were analyzed via a 5 × 9 × 9 analysis of variance having repeated measures over the second two dimensions.

*NS = not significant

**Partitions of the variance are presented in Tables 7 and 8.

graphs, and Cartoons followed in that order. Readers agreed that Exemplary Teaching Practices were needed more than any other section (Table 10). Reading Research Review, ECRI-sponsored Projects, Reading Keys, Feature Article, Photographs, Letters to the Editor, Cartoons, and About the Author followed in that order. Readers agreed that the section having least value to them was Letters to the Editor (Table 11). Photographs, Cartoons, Reading Keys,

Table 7

Inter-section Comparison[a] of all Ordered[b] Pairs of NEWSLETTER *Sections*

(Difference Matrix[c])

	C	*F*	*A*	*G*	*E*	*D*	*I*	*H*	*B*
C		3	20*	65**	71**	181**	189**	193**	221**
F			17*	62**	68**	178**	186**	190**	208**
A				45**	51**	161**	169**	173**	191**
G					6	116**	124**	128**	146**
E						110**	118**	122**	140**
D							8	12	30**
I								4	22**
H									18**
B									

[a]The variance was partitioned via the Newman-Keuls Sequential Range Statistic.

[b]NEWSLETTER sections are ranked on both the abscissa and the ordinate in the ascending order of total rating scores.

[c]Any score in the matrix is the absolute difference between the total rating scores assigned to the sections directly opposite it on both the abscissa and the ordinate. The lowest rating indicates highest desirability.

*Denotes significance beyond the .05 level.

**Denotes significance beyond the .01 level.

About the Author, ECRI-sponsored Projects, Reading Research Review and Exemplary Teaching Practices (tied ranks), and the Feature Article followed in order of ascending value.

One of the most significant findings of this evaluation emerged from complementary results produced by two independent analyses: (a) the degree to which the *Newsletter*'s contents reflected its objectives (an analysis of Item III data), and (b) the degree to which the *News-*

Table 8

Inter-category Comparison[a] of all Ordered[b] Pairs of Rating Categories

(Difference Matrix[c])

	12	*11*	*10*	*9*	*8*	*5*	*4*	*6*	*7*
12		36**	49**	59**	64**	64**	68**	70**	82**
11			13**	23**	28**	28**	32**	34**	46**
10				10**	15**	15**	19**	21**	33**
9					5	5	9*	11*	23**
8						0	4	6	18**
5							4	6	18**
4								2	14**
6									12**
7									

[a]The variance was partitioned via the Newman-Keuls Sequential Range Statistic.

[b]Rating categories are ranked on both the abscissa and the ordinate in the ascending order of total rating scores.

[c]Any score in the matrix is the absolute difference between the total rating scores assigned to the rating categories directly opposite it on both the abscissa and the ordinate. The lowest rating indicates highest desirability.

*Denotes significance beyond the .05 level.

**Denotes significance beyond the .01 level.

letter's contents fulfilled its readers' needs (an analysis of Item II data).

Item II data show that readers read most often, requested most often, and valued highest all *Newsletter* sections relating to exemplary classroom practices (Sections A, C, F, and G; Tables 9, 10, 11). Item III data show that readers consistently rated these same sections higher than the rest (Table 7). The fact that, among the rating categories, the *Newsletter* was rated lowest on practicality and applicability (Table 8) can be explained by the significant nonadditive variance (Table 6; SECTIONS by RATINGS) that remained beyond significant SECTIONS and RATINGS main effects. A further analysis of these data shows that Exemplary Teaching Practices, as well as the other sections carrying exemplary teaching practices, were exempted from the low practicality and applicability ratings. Partitioning the nonadditive variance placed Reading Research Review, Exemplary Teaching Practices, Reading Keys, and the Feature Article in a group statistically above the remaining sections insofar as practicality and applicability were concerned.

Table 9

Ranking of* NEWSLETTER *Sections According to Readership Strength*

Readership Groups	NEWSLETTER *Sections*								
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
ET	1	6	2	8	5	3	4	7	9
EA	1	6	2	7	4	3	5	8.5	8.5
STA	1.5	5	3	6.5	4	1.5	8.5	8.5	6.5
CTA	1	4	2.5	7	5.5	2.5	8.5	8.5	5.5
O	1.5	5	1.5	7	4	3	6	8	9
TOTAL*	6.0	26.0	11.0	35.5	22.5	13.0	32.0	40.5	38.5

*Ranks were analyzed via the Friedman two-way analysis of variance, where:

$\chi^2_r = 33.94$ and: $p < .001$

*The lowest total score identifies the most heavily read section. A χ^2_r of 33.94 indicates significant differences among all pairs of rankings, i.e., section A is read significantly more than its closest competitor, section C, and so on.

Table 10

Ranking[a] of NEWSLETTER *Sections According to Readership Need*

Readership Groups	NEWSLETTER *Sections* *A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
ET	5.5	9	1	5.5	4	2	3	8	7
EA	5	8	1	6	3.5	2	3.5	7	9
STA	5	5	1	9	3	2	7.5	7.5	5
CTA	3	8	2	8	4	1	5	8	6
O	5	9	1	8	3	2	4	6.5	6.5
TOTAL*	23.5	39.0	6.0	36.5	17.5	9.0	23.0	37.0	33.5

[a]Ranks were analyzed via the Friedman two-way analysis of variance, where:
$\chi_r^2 = 33.94$ and: $p < .001$

*The lowest total score identifies the most needed section. A χ_r^2 of 28.07 indicates significant differences among all pairs of rankings, i.e., section C is significantly more valuable than its closest competitor, section F, and so on.

BASIC RESEARCH

While the Exemplary Reading Center's mission is primarily that of program development, training, and dissemination, it has some commitment to basic research, which is part of the total evaluation program. In-service programs involve instructional strategy. Basic research has been supported because of its focus on basic psychological principles underlying some of the instructional methods used in our in-service programs. Typically, basic research is sponsored by the Reading Center in cooperation with other agencies such as the University of Utah. Two doctoral dissertations* were carried out

*Alter, Madge. Identification of high probability responses and their use as reinforcers. Doctoral dissertation, University of Utah, 1968.
Chan, Adrian. An analysis of Premack's rate differential response theory. Doctoral dissertation, University of Utah, 1968.

under the direction of Drs. Della-Piana and Sloane. Both studies dealt with Premack's theory that of two events (A and B) the one occurring more frequently (A) will reinforce the lower-frequency event (B).

Dr. Alter was concerned with developing procedures for identifying high-probability responses and determining their stability and their utility as reinforcers for low-frequency responses. Premack hypothesizes that if a response occurring at a higher rate is made contingent upon a lower-rate response, the high-rate response can be used to reinforce (increase the frequency of) the low-rate response. If this principle is to find practical application in the classroom, a procedure must be devised whereby teachers can chart response frequencies for commonly occurring classroom activities.

An initial study was designed to do just that. A method was developed for identifying high- and low-frequency activities for individual pupils. Commonly occurring classroom activities were

Table 11

Ranking[a] of NEWSLETTER *Sections According to Perceived Value*

Readership Groups	NEWSLETTER *Sections*								
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
ET	9	4	7.5	2	5.5	7.5	5.5	3	1
EA	7.5	4.5	7.5	2	7.5	7.5	4.5	3	1
STA	7	3	7	3	7	7	7	1	3
CTA	8.5	7	5	1.5	5	8.5	1.5	3	5
O	8	5	6	2.5	8	8	4	2.5	1
TOTAL*	40.0	23.5	33.0	11.0	33.0	38.5	22.5	12.5	11.0

[a]Ranks were analyzed via the Friedman two-way analysis of variance, where:

$\chi_r^2 = 24.64$ and: $p < .01$

*The lowest total score identifies the least valuable section. A χ_r^2 of 24.64 indicates significant differences among all pairs of rankings, i.e., sections D and I are significantly less valuable than their closest competitor, section H, and so on.

paired and presented to the children who were to choose between the alternatives in each pair. The activities were ranked according to attractiveness. The reliability or stability of the rankings was assessed using a test-retest procedure. The validity of the rankings was determined by a correlational analysis between paired-comparison rankings and actual frequency counts of the same classroom behavior as that used in the paired-comparison presentations, and by a correlation of paired-comparison rankings with rankings on a two-choice task using an apparatus which presented reading or arithmetic materials at the press of a button.

The activity categories were obtained by observing frequently and regularly occurring classroom events. Simple line drawings of each activity were made. The drawings were arranged in all 21 possible pairs, and slides were made of each pair. Slides were shown while a synchronized taped voice asked: "Which of these activities do you do? *Special Activities* (like drawing maps, coloring, or cutting out decorations) or *Checking With The Teacher* (to see whether an answer is right, to find out the assignment or to tell him something interesting)."

The four highest ranking activities for males were *Arithmetic*, *Reading*, *Special Activities*, and *English*, in that order; for females they were *English*, *Special Activities*, *Reading*, and *Arithmetic*, in that order. Stability of highest and lowest paired-comparison choices for a two-week interval was determined for 45 third graders. Agreement was determined as follows: An activity which ranked 1 (highest frequency of the seven activities) on the first administration of the paired comparison task was counted as an agreement in choice two weeks later only if the activity was chosen with sufficient frequency to place it between the ranks of 1 and 3.5. If the activity was ranked 7 on the first administration, agreement two weeks later meant ranking from 3.6 to 7 on the retest. Seventy-six percent of the originally high frequency responses met the criterion; 96 percent of the low-frequency responses met the criterion. Thus, the stability rankings of extreme cases were adequate, particularly for initially low-frequency responses.

Two concurrent validation methods were explored to determine whether the paired-comparison rankings were similar to those obtained by other techniques with apparently greater face validity. The first involved tallying the frequency of the seven activities within a classroom using an Esterline-Angus 20-pen Event Recorder adapted for recording frequency and duration of responses. No significant relationships were found between choices on the paired-comparison

presentation test and actual behavior within the classroom during the free-choice period. The second validation method employed was a correlation of the paired-comparison frequency rankings of activities with the rankings based on a two-choice task. The two-choice task was composed of reading and arithmetic materials. The reading materials were short paragraphs from an SRA *Reading Laboratory* modified to obtain similar duration of response for each selection. Arithmetic materials contained addition and subtraction problems from the ABC *Modern Mathematics Series*, Grade 1. Agreement for reading activity between the two-choice test frequencies and paired-comparison frequencies was 80 percent for males and 92 percent for females. Agreement for the arithmetic activity was much lower. Thus, a simple paired-comparison approach to getting frequency rankings was highly predictive of rankings based on a two-choice task using an apparatus which allowed a choice between arithmetic and reading problems.

The final stage of Dr. Alter's study followed the Premack paradigm. Each child in the experimental group participated in three sessions. The first was a baseline session to determine the child's high-frequency response (arithmetic or reading). The second was a contingency session in which the high-frequency response (arithmetic or reading) could be performed only following performance of the low-frequency response. The third (extinction) session was a return to baseline conditions in which there were no contingency relationships established. A control group also participated in three sessions, which were conducted under baseline or noncontingency conditions.

Subjects were 24 third graders (12 male and 12 female). The design was a 2 (sex) × 2 (experimental-control group) × 2 (high probability reading-high probability arithmetic) × 3 (sessions) factorial with repeated measures on sessions. Each subject had 40 trials within a session. The apparatus used for presenting materials was the two-choice task apparatus referred to above. Under baseline conditions both response buttons were operative and produced stimulus materials (arithmetic or reading) whenever they were pressed. During the contingency session one of the two response buttons was inoperative until the other button was depressed, thus forcing the high-frequency activity to be contingent upon performance of the low-frequency activity.

Major Findings

The major findings of this study were that: Baseline performances were highly stable (control group performance did not differ significantly across sessions I, II, and III); experimental and control groups did not differ significantly under session I baseline conditions, nor did they differ significantly in session III during which both groups were tested again under baseline conditions; low-probability response frequency for experimental subjects was significantly higher in session II than in session I and was higher for the experimental group than for the control group, and the results were the same whether the high-probability response was reading or arithmetic.

Thus, a simple paired-comparison procedure for determining response probabilities was developed. Frequency rankings of activities were found to be highly stable over a two-week period for high- and low-frequency activities. Validity of the paired-comparison rankings was supported by high correlation with frequency of a choice in the two-choice task. Validity of paired-comparison rankings was also supported by an increase in initially low-probability responses produced under conditions in which they were requisites to performing high-probability activities.

Chan's Study

Dr. Chan's study was an outgrowth of Dr. Alter's investigation. While Dr. Alter's work supported Premack's earlier findings, there remained the question of the extent to which the reinforcement effect of high-probability responses was due to *frequency of reinforcement or response rate*. Three studies were conducted to answer this question. Experiment 1 manipulated response rate, while holding reinforcement frequency constant, to evaluate the role of rate alone. Experiment 2 manipulated reinforcement frequency, while holding response rate constant, to evaluate the role of reinforcement frequency alone. Experiment 3 varied both reinforcement frequency and response rate to evaluate the role of both factors simultaneously.

The results of all three experiments suggest that the instructional variable became a contaminating factor. When the experimenter made comments such as "Go faster on this button to get to the side you like," the results clearly yielded rate changes (increases) as a function

of reinforcement frequency and not response rate. But, for minimal cues given to the subjects, no rate change occurred as a function of changes in response rate or reinforcement frequency. Thus, the role of the instructional variable needs to be explored further before unequivocal interpretations can be made of the relative role of *response rate* and *reinforcement frequency* in findings supporting Premack's hypothesis.

APPENDIX A

Scoring Summary

CSE	*Contingent Stimuli Established*
S^{rp}	Scored if the teacher offers and describes a reward for appropriate behavior
S^{rn} *avoid*	Scored if the teacher describes a punishment that will be imposed for inappropriate behavior
S^{rn} *escape*	Scored if the teacher promises to allow his pupils to escape a promised punishment if they behave appropriately
CSA	*Contingent Stimuli Applied*
S^{rp}	Scored if the teacher rewards his pupils as promised
S^{rn} *avoid*	Scored if the teacher punishes his pupils as promised
S^{rn} *escape*	Scored if the teacher allows escape from a punishment he has imposed
RCSA	*Response Contingent Stimuli Applied*
S^{rp}	Scored if the teacher verbally rewards his pupils without first promising reward
ext	Scored if the teacher rewards his pupils with extrinsic reinforcers
tok	Scored if the teacher rewards his pupils with a token (anything eventually traded for a reward)
S^{rn}	Scored if the teacher punishes his pupils without first warning them
T.O.	Scored if the teacher punishes his pupils with time out (isolation)
Comment Spaces:	Provided for observer's comments
Timing:	Used to record observation beginning and ending times
Scoring Responses:	Each time a category is scored, the time the behavior occurred is noted in the proper subcategory spaces

APPENDIX B

Time Begin *2:10* Time End *2:40*

Sample Observation Record

CSE — S^{rp} — COMMENT
2:14, 2:21, 2:24

S^{rn} — COMMENT
AV
2:23

S^{rn} — COMMENT
ES

CSA — S^{rp} — COMMENT
2:31

S^{rn} — COMMENT
AV

S^{rn} — COMMENT
ES

RCSA — S^{rp} — COMMENT
2:39

tok ext

S^{rn} — COMMENT

T.O.

GENERAL COMMENT:

APPENDIX C

General Rating-scale Procedure

Certain teacher responses are listed on the rating scale and are rated according to the code and procedure outlined below. Raters need two sharpened pencils, a clipboard with a stopwatch attached, blank note paper, a rating pack, and a supply of rating sheets.

Rating sheets are divided into nine 30-second intervals on the horizontal axis and five teacher response categories on the vertical axis. Rating packs are made up of pictures of 10 children within a given classroom. The children's names are printed on their pictures.

General Coding Procedure

1. Ratings are to be coded only when the regular teacher is present.
2. Raters are to draw a picture from the top of the shuffled, face-down stack and record or code that child's behavior and the teacher's responses to this child for 4½ minutes, select another, observe for 4½ minutes, and so on until all 10 children have been observed.
3. Raters are to observe during the first 10 seconds of each of the nine 30-second rating periods, within the 4½ minutes for each child, noting which behavior occurs.
4. Raters are to record or code the observed behavior during the final 20 seconds of each 30-second observation interval. They are not to observe during this time.
5. Raters are to record or code all behavior that occurs during an observation interval.
6. If a certain element of behavior occurs more than once during an observation interval, raters are to record or code all of the observations which were noted, unless it is indicated otherwise in the instructions.
7. Raters are not to respond to any child in the classroom but are to ignore the children.
8. Raters are to use only the coding criteria as outlined in the instructions. If an element of behavior cannot be rated according to instructional criteria, note that it cannot be. Do not try to judge behavior or its intent.
9. Raters should be trained in the use of the scale according to the detailed procedures on the training sheet before attempting any data collection.

Student Behavior

Student behavior (SB) is listed as Area 1 and is located on Row 1 of the rating scale.

Rating Procedure

1. Student behavior is to be rated within every 30-second period.
2. The ratings are to be based upon a 4½-minute sample of the target child's behavior. Target children are to be rotated every 4½ minutes.
3. Student behavior is to be listed as either desirable (D) or undesirable (U). If no undesirable behavior occurs during a 30-second rating interval, the interval is coded D. If one or more instances of undesirable behavior occur, that interval is coded U.

Undesirable Behavior Includes:

1. talking aloud without permission;
2. making nonverbal noise such as tapping a pencil on a desk;
3. wandering around the room without instruction from the teacher;
4. disruptive motor behavior such as fighting, wiggling, and poking other children, even if the behavior is instigated by another child;
5. slowly or improperly getting or returning materials;
6. failing to begin, continue, and complete classwork on time as directed;
7. failing to attend during teacher presentations; and
8. leaving the seat without permission, unless regularly permitted to do so.

Nonverbal Teacher Behavior Toward Target Child

Nonverbal teacher behavior toward the target child (NV-T) is listed as Area 2 and is located on Row 2 of the rating scale.

Rating Procedure

1. The rater is to code any nonverbal teacher behavior toward the target child:
   p—if the teacher points at the child;
   c—if the teacher touches or otherwise contacts the child;

h—if the teacher reacts by smiling, winking, nodding, sticking tongue out, frowning, grimacing, head shaking, looking, or any response given with the head; and

a—if the teacher approaches the child, touches his desk or materials thereon, but does not touch him.

2. In addition, a plus sign (+) is added to any of the above codings when the teacher's behavior is unquestionably approving, and a minus sign (—) when the teacher's behavior is unquestionably disapproving. Disregard the plus (+) and minus (—) signs if in doubt.

Verbal Teacher Behavior Toward Target Child

Verbal teacher behavior toward the target child (V-T) is listed as Area 3 and located on Row 3 of the rating scale.

Rating Procedure

1. The rater is to code any verbal teacher behavior toward the target child:

(+)—if the teacher states that the target child is engaging in a D behavior, is not engaging in a U behavior, is to receive some positive reinforcement, or otherwise praises him;

(—)—if the teacher states that the child is engaging in a U behavior, is not engaging in a D behavior, is to receive something aversive, or otherwise reprimands or criticizes him;

I—if the teacher gives an assignment, answers a child's question, indicates what the child is to do or how he is to do it, or otherwise instructs him;

O—if the teacher verbally interacts with the child in a way not clearly part of another code.

2. If the teacher specifies another child or children *along with* the target child, rate the teacher's verbal behavior in row V-T. The target child may be the only child spoken to or he may be specified by name along with other children. A rating is not made in row V-T if the teacher does not in some way specify the target child as his spoken target while excluding most of the others in the class.
3. Note that in row V-T more than one code can often be recorded. For instance, if the teacher instructs the target child and praises

him in addition, the rating becomes I+. An example of this would be, "Johnny, when you finish reading, you may go to recess." A + only code does not include an instruction; e.g., "Johnny, you've worked so hard today that you may go to recess early." Coding combinations are similarly used with the minus sign.

Verbal Teacher Behavior Toward Other Than Target Child

Verbal teacher behavior toward others (V-O) is listed as Area 4 and is located on Row 4 of the rating scale.

Rating Procedure

The rater is to code verbal teacher behavior which is in no way directed towards the target child. The rater should note whether the teacher specifies another child or children or whether she directs her statement to the class in general.

The codes and procedures used for Area V-T are also applicable in Area V-O.

General Character of Teacher Interaction

General character of teacher interaction (TI) is listed as Area 5 and is located on Row 5 of the rating scale. Interactions may include questions, statements, explanations, prompts, probes, calling on a child, etc.; and, depending upon the teacher's intent, any of these can be academic instruction, schedule instruction, or behavior management.

Rating Procedure

1. The rater is to code at least one type of teacher interaction within every rating interval:
   i—if the teacher interacts with one student;
   g—if the teacher interacts with a group of students ranging from two children to approximately one-half of the class; and,
   c—if the teacher interacts with more than one-half of the class.
2. The teacher may work with a single individual as well as speak to the class during a single 10-second interval; therefore, there is a

good possibility that all three codes may be used during any one rating interval.

3. Additional code specifications are used in conjunction with the above when:
   (a) The interaction is basically academic. If the interaction concerns academic work or content, circle the i, g, or c. Examples are: "6 + 8 are 14," "Your answer is correct," or "I am sure you remember who saw the bunny."
   (b) The interaction basically concerns scheduling. If the interaction concerns changing or moving activities, locations, materials, etc., as a function of the class schedule, prime the i, g, or c. Examples are: "Put your papers away now," "The time is nearly up," or "Let's sit at the large table."
   (c) The interaction basically concerns behavior management. If an interaction is an attempt by the teacher to get a child to stop emitting a U behavior or an attempt to get a child to emit a D behavior, underline it. Examples are "Turn around in your seats," "Be quiet."
   (d) If an interaction cannot be coded as academic, scheduling, or behavior management, code it i, g, or c.

Summary

AREA 1 *Student Behavior*

U (undesirable) or D (desirable), (score one—give U preference)

AREA 2 *Nonverbal Teacher Behavior Toward Target Child*

p (points), c (contact), h (head), a (approaches), (score all) Score + or — if appropriate

AREA 3 *Verbal Teacher Behavior Toward Target Child*
(score applicable behavior)

+ positive
— negative
I instruction
O other verbalization

AREA 4 *Verbal Teacher Behavior Toward Other than Target Child*
(score all)

+ positive
— negative
I instruction
O other verbalization

AREA 5 *General Character of Teacher Interaction*
(score applicable behavior)

i individual academic interaction
g group academic interaction (less than half the class)
c class academic interaction

i′ individual schedule instruction
g′ group schedule instruction (less than half the class)
c′ class schedule instruction

i individual behavior management
g group behavior management (less than half the class)
c class behavior management

i individual interaction, other
g group interaction, other
c class interaction, other

APPENDIX D

PLEASE READ THE DIRECTIONS CAREFULLY BEFORE ATTEMPTING TO COMPLETE AN ITEM.

I

Please check (✓) the appropriate boxes

	Yes	No
Do you read the ECRI Newsletter?	☐	☐
Would you like to continue receiving future issues?	☐	☐

If you answer No, we must have your name and address so that we can delete it from our Newsletter mailing list. See item VI. below.

III.

Please rate each Newsletter Section (rows A through I) on each Rating Category (columns 4 through 12) Using the five point scale, fill in each box with the number which best represents your view of how clear, informative, important, etc , each Newsletter Section is:

1	2	3	4	5
excellent	above average	average	below average	poor

II

Please check (✓) the appropriate boxes in Columns 1 through 3.

NEWSLETTER SECTIONS	Sections which you read regularly	Sections which you would give more space in the Newsletter	Sections which are of no value to you	RATING CATEGORIES	Influential	New, original, innovative	Practical, sound, realistic	Applicable	Useful	Important, significant, valuable	Interesting, challenging	Informative, instructive	Clear, easily understood	RATING CATEGORIES
A Feature article (front page)														
B About the author														
C Exemplary teaching practices														
D Letters to the editor														
E ECRI sponsored projects														
F Reading research review														
G Reading keys														
H Cartoons														
I Photographs														
	1	2	3		4	5	6	7	8	9	10	11	12	

IV.

Please check (✓) the box which most accurately represents your current job category. If you have to use a box marked **Other**, please specify your current position in the appropriate box.

	Teacher	Administrator	Other
Elementary			
Secondary			
Jr. College			
College			
Other			

V

Comments: _____

VI.

Name and address of evaluator (optional unless answer to question two, item I is No):

Name _____ Address _____ Zip _____

Evaluating a National Program: The Training of Teachers of Teachers

BERTRAM B. MASIA and P. DAVID MITCHELL
Case Western Reserve University

In this paper we shall attempt to identify problems and issues that arise in the evaluation of a national project. Because our evaluation studies are still in process, we shall not report results.

Our specific point of reference is the Project for the Training of Teachers of Teachers, which was recently supported by the United States Office of Education under Title XI of the National Defense Education Act and more popularly referred to by its designers as the "Triple-T" Project.

We use the word "national" to refer to the scope of the project. Its setting was the vastness of the country and its scenario required many hundreds of educators and laymen to perform in many places studying, learning, discussing, disagreeing, creating, planning, writing, and rewriting.

We also use the word "project," rather than "program," to indicate that it was a relatively short-term affair, a one-time design which may or may not be repeated in its original or revised form. In other words, to the extent that it was considered an experiment, a trial effort along certain lines, we choose to call it a project.

The evaluation of this material project consisted of appraising the performance characteristics of the project design as it was implemented over a nine-month period.

To satisfy the curiosity of the uninitiated, let me deal with the meaning of the project title. From the very inception of the project this title, the Training of Teachers of Teachers, seemed to many to be a riddle it was never intended to be. It caused too many participants, even entire groups of participants, to stray from the main path often or, in several instances, never to travel that path at all. Perhaps

the title is ambiguous. If it was difficult to decode, there were many statements in the form of briefings and commissioned papers (2, 4, 6) that gave rather unambiguous definitions. B. O. Smith (7) has defined the phrase by delineating four fairly distinct levels or strata of instructional personnel. It is the education and training of persons at level four that represented the focus of the Triple-T Project. Smith writes:

> The first level consists of teachers who man the classrooms of the elementary and secondary schools, and I suppose one must these days include preschools and junior colleges and colleges as well. The second level consists of supervisors, directors of instruction, and so on, who work with the personnel at the first level. The third level consists of college teachers who man the classrooms and laboratories in which persons who occupy the first level are prepared to do their work. These latter persons we refer to as teacher trainers, teacher educators, or teachers of teachers, depending on one's semantic taste. Level four is comprised of the college teachers who man the classrooms and laboratories in which the personnel of levels two and three are trained. These are the persons who are responsible for the preparation of the teachers of teachers.

Bigelow (2) put it in a somewhat different form:

> I believe I can say with reasonable assurance, that hardly anyone ever jumped out of bed in the morning and said with gusto, 'Bully for me; I'm a trainer of teachers of teachers, and I am going to work today.' However, there are a host of people who might, if they chose, say precisely that. Besides university professors in schools of education and liberal arts, many people are actively engaged in the task of training teachers who teach teachers. For example, there are librarians, educational media specialists, critic teachers, school administrators of every rank . . .

There may be less than perfect reliability among expert sorters in assigning educational roles to Smith's levels. But on one crucial matter there is little room for disagreement: The training of teachers of teachers for educational institutions of all types takes place in the graduate schools of our universities and takes place throughout the graduate school.

That there is much room for its improvement and that it should not be confused with the training of researchers are two of the key ideas of the Triple-T Project.

Perhaps these distinctions can be made concrete if I describe how the project came to be. In tracing the roots of the Triple-T Project, the educational rubric that is the sole focus of our attention is teacher training. More specifically, the project has to do with the formulation of strategies, at the local level, for training teachers. This is a matter that affects all of us directly or indirectly because most of us are or have been teachers and are forced to wrestle with pedagogical matters and because the public criticism that hurts us most as educationists is that which has to do with the manner in which we prepare teachers.

The direct lineage of the Triple-T Project can be traced to the summer institutes for high school teachers of mathematics and the sciences supported by the National Science Foundation, and the programs for foreign-language teachers supported by the Office of Education.

Sputnick, of course, was the precipitating cause. In later amendments to the National Defense Education Act, the Office of Education was authorized to expand its institute program to other subjects and matters such as history geography, economics, civics, English, English as a foreign language, disadvantaged youth, educational media, reading, and so forth. Institutes for elementary teachers were also added in most of these domains of education.

While the summer institute program continued on campuses throughout the country, variations developed in the form of part-time academic-year institutes and other forms of in-service training. This, in turn, led to the Experienced Teacher Fellowship Program in which teachers are theoretically in residence on campus as mature teachers, not doctoral candidates.

The Prospective Teacher Fellowship Program was designed with the recognition that reeducation or remediation had become the exclusive focus of funding. Essentially this program is a first cousin to the Master of Arts in Teaching programs that had been supported by the Ford Foundation since the late 1950s in the graduate schools of quite a few private and public universities and liberal arts colleges. To a large extent, however, the Prospective Teacher Fellowship Program replaced private foundation funds with money from Washington. The Prospective Teacher Fellowship Program provided the door by which the federal government entered the preservice sector of teacher training.

The direct involvement of the federal government in teacher train-

ing is relatively recent. But the pace of its activity and the degree of its interest in this matter have been accelerating rather sharply during the past few years. As a result of monitoring of these programs by federal officials, reports from interested individuals in higher education and in the schools, formal evaluation studies, and much debate and discussion inside and outside the Office of Education, a number of facts became clear. First, that the money managers in Washington felt that they weren't getting enough return on their investment solely on the basis of the relatively few teachers reached in these programs. Each year only one percent of all teachers in the country were reached directly in all programs supported by the Office of Education (3). Coupled with this was the not-surprising finding that the teachers receiving training had, for a variety of reasons, little or no impact on their schools, departments, and colleagues in terms of sharing and communicating what they had achieved and mastered in their training. The designers of the programs were either naive or overoptimistic as to the impact of the training programs on educational personnel back home. At best, the program participant himself had changed in desired directions, but he was either unwilling or incapable or not allowed to spread the gospel.

Even more disquieting was the news that nothing was changing on campus. The typical teacher-training program focused on knowledge and understanding in the discipline, which was good in itself but gave only lip service to what had been considered crucial—namely, the transformation of discipline content into pedagogical skill.

Finally, the programs were restricted to institutions of higher education. Participation in a systematic fashion in local program design and execution by other educational institutions was difficult to find.

The Tri-University Project in Elementary Education and the creation in 1966 of nine institutes for college teachers, the so-called trainers of teachers, represent a second generation of training-program designs. Previous programs to improve teaching skills reached a small percentage of teachers. One problem was how to use limited funds to reach more teachers so as to have maximum impact for the time, effort, and money invested. A possible solution to this problem was to go for the "gatekeepers," the teachers of teachers; a solution to bridging the chasm on campus between pedagogical and nonpedagogical departments was the idea, in Olson's (5) words, "of bringing together in a 'double practicum' teachers of teachers, teachers, and

elementary school children in a single project to improve the quality of college programs for training teachers and consequently the quality of teachers."

So much for the teacher-education roots of the Triple-T Project. Two other types of roots need to be mentioned in order to fully explain the project design. The first has to do with the manner in which proposals had been requested for teacher education programs. Generally they were prepared with an explicit set of guidelines at hand and with a relatively brief period for writing. Dissatisfaction with this arrangement was widespread in the Office of Education, among proposal reade and in the universities. It was generally agreed that less prescription and less structure were needed, along with more time for project development. Categorical funding was still the order of the day, but more degrees of freedom were being allowed.

The third and final root of the Triple-T Project I shall call civil rights and urban education. Just as there was general agreement about an alienation between the school of education and the rest of the university with respect to the training of teachers, there was also concern about an alienation between the university and the community including the common schools and the people they serve (particularly poor people). Stated in another manner, the belief had developed that a key to the problem of training teachers was that the academic faculty, the pedagogical faculty, precollegiate school people, and laymen should learn to talk to each other as equals—the principle of parity—so that the role of each in teacher training could be made more explicit.

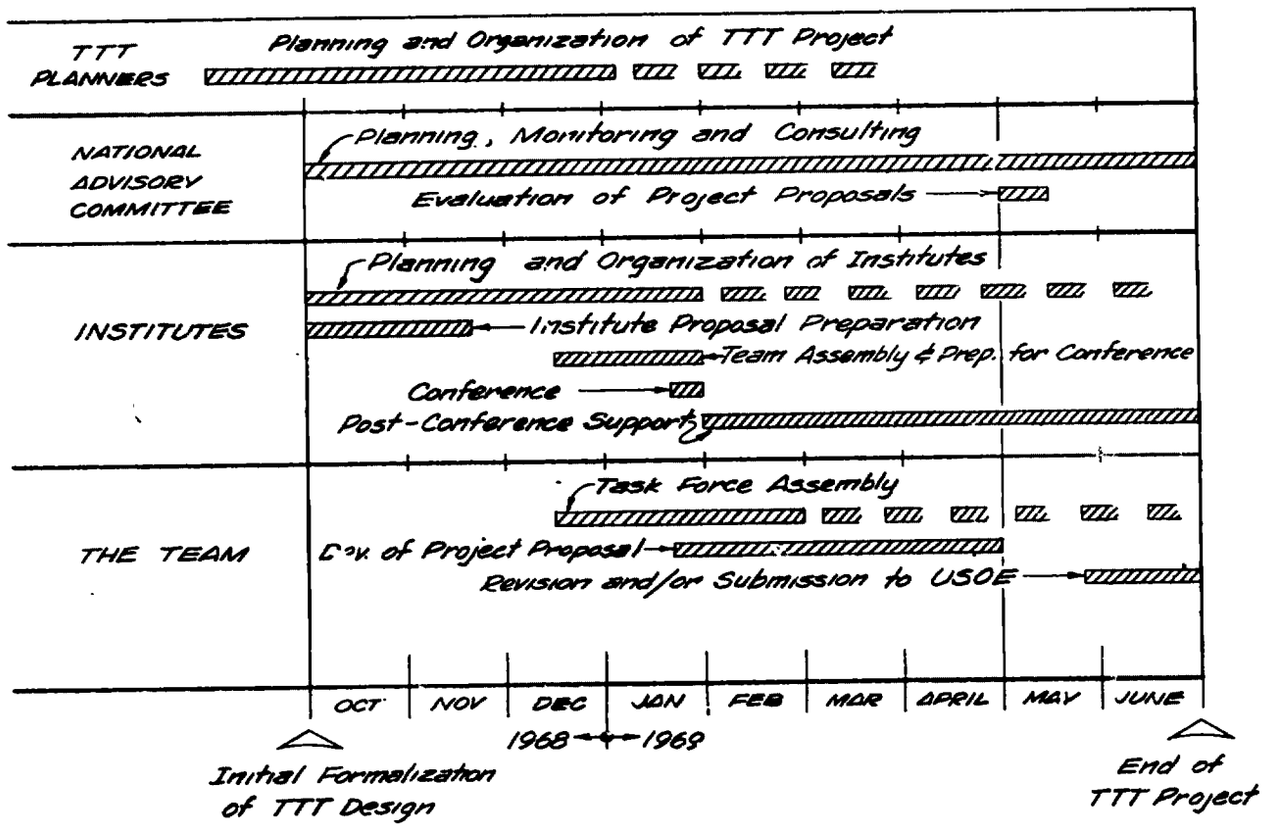
The Triple-T Project could, therefore, be viewed as the fostering and nurturing of interactions among these interests toward the development of local training programs, supported with federal funds, in which these several interests would share the responsibility, in accordance with their public roles, in the training of teachers of teachers.

The Design of the Project

The design of the project was clear-cut. It can be viewed in two ways—temporally and structurally. The time line in Figure 1 shows a total duration of the project, at least in a formal sense, of approximately nine months. This period can be divided into three overlapping sub-

Figure 1

The TTT Project—Development and Execution Activities



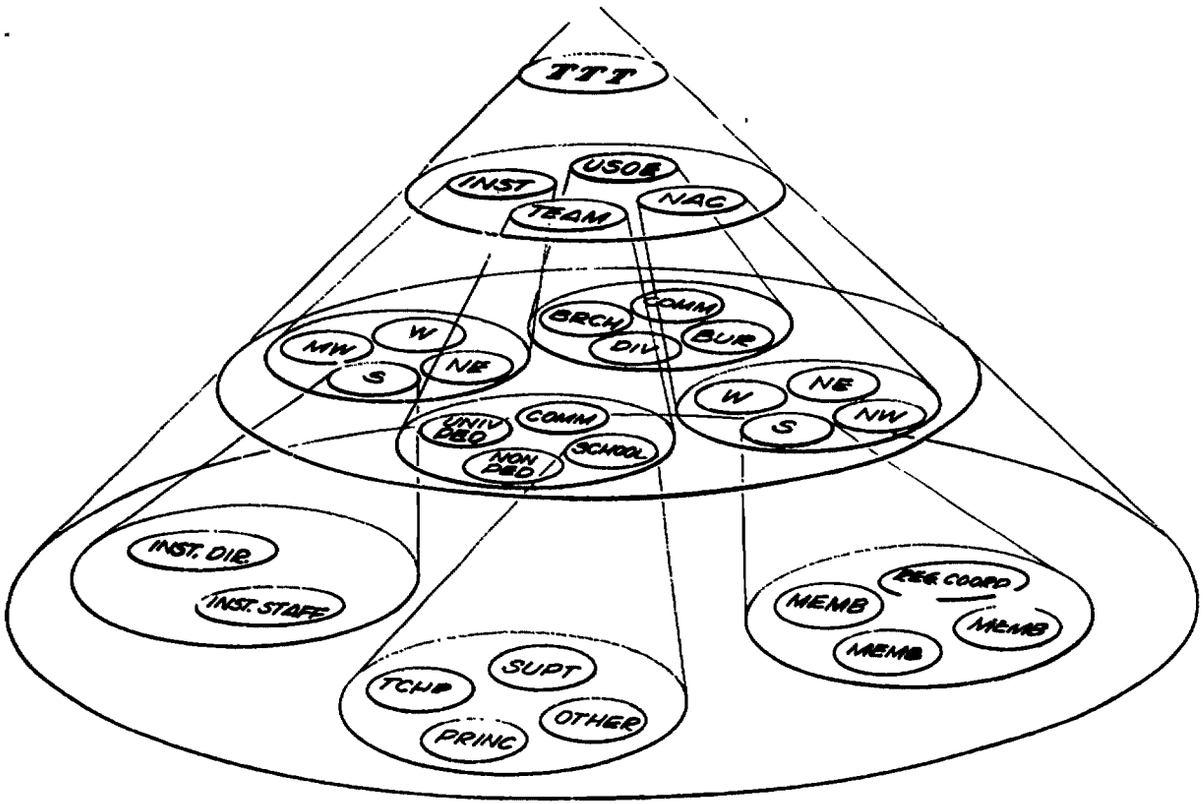
periods: 1) a subperiod of preparation, from October through January, requiring the selection and design of four regional institutes to serve as catalytic, training, and resource agents, and the selection of some 65 places at which Triple-T teams would be assembled; 2) a subperiod (mid-November through March) of team assembly and the search by the team or task force for a Triple-T identity; and 3) a subperiod of task force development of a project for the training of teachers of teachers, the appraisal of this project by a National Advisory Committee; and finally, revision or reformulation of the project in the light of the National Advisory Committee review and submission of the document as a proposal under the new Educational Personnel Development Act. This took place from early January to July 1.

In Figure 2, there are four types of components in the design structure: the Office of Education, a National Advisory Committee, the regional institute, and the team or task force. These four components were in more or less continuous interaction with each other throughout the duration of the project. The functions of each component were made clear at the outset. A major task of evaluation was to determine whether such functions were carried out and if not, why not. In brief, we were evaluating the performance of the design over time.

Figure 2 indicates that components were also broken down as subcomponents, and could be broken down even further, generally for purposes of data analysis. Office of Education officials who played specific roles in the project or who participated in a less formal and unplanned—but certainly nonrandom—manner came from different levels of that hierarchy—branch, division, bureau. A subcomponent could also be viewed as a document representing an official's position. For example, the project had to deal with a non-Triple-T pronouncement by the Commissioner of Education who stated that it was the policy of the Office to allocate a significant portion of the funding of all Office of Education programs to the urban sector.

The subcomponents of the National Advisory Committee—the project management, so to speak,—were four groupings of the Committee which were organized regionally for the purpose of more effectively relating to the four regional institutes at Hunter, Georgia, Michigan State, and UCLA, and to the task forces in each region. Since these subcomponents were established after the project began, they required a significant increment in the size of the National Advisory Committee (almost doubling it from 12 to 22) and a significant change in its representation (the expansion allowed for more

Figure 2
Hierarchy of Analytical Levels



equitable representation of nonuniversity people). The consequences of this decision were felt throughout the structure, causing much strain and noise and requiring additional effects to be theorized and monitored by the evaluation staff.

The task forces component could be viewed as a set of 64, as four regional subsets, or as 64 separate and idiosyncratic phenomena. Furthermore, the project design required that a task force be assembled on the basis of two principles, namely parity and verticality. Thus, members of task forces were two hats, one representing an educational sector such as university nonpedagogue, and the other representing an organizational level such as teacher or professor on the one hand and dean of arts and sciences or superintendent of schools on the other. The task force structural property of verticality is based essentially on the following argument by Haubrich (4):

> What seems to come through on the socio-psychological end of the matter in combination with the administrative theory as postulated by Griffith, is that innovation from the bottom is virtually impossible in the educational system and that the independence of sub-systems within the organization isolates each group from change activity. Clearly, the question that we face in organizational and bureaucratic change is linking the functionaries, providing for communication between those at the top and at the bottom, engaging in programmatic activities which are centered in the situation where the functionaries are at work, and lastly, the selection of school systems and individuals who seem to have a propensity for testing out new ideas.

With regard to the regional institutes, which had essentially educational and consultative functions, Figure 2 indicates that the evaluation can focus upon several different levels. When we deal with ideas about institutes in general, such as their educational function, we treat the institutes as a set of four. When we deal with ideas that are unique to an institute or to a region, such as the emphasis on creativity training at the Georgia institute, we focus on a specific institute. However, data obtained at one level can be used in the study of a question belonging to a more general level of analysis. Therefore, at the general level of institute analysis, the data concerning creativity training at Georgia that we had collected can be utilized in a study of institute program design.

Thus, there are several levels at which the project may be described, and these are shown in Figure 2. Each distinguishable feature at one level may represent a wealth of detail when examined on a larger scale

at a lower level (1). As we progress from the larger, more general system, we come progressively closer to the level where the behavior of every project participant can be identified and described. When moving down the hierarchy of levels, as part of our search for an understanding and an appraisal of the processes by which 64 projects were produced and evaluated, our task becomes ever more complicated. While it was clear from the very beginning of the project that we could not monitor every interaction between individuals, or even between groups, it was necessary to work simultaneously at several levels. Thus, while we focused on the level of the four components or subsystems, we learned which areas of activities at lower and more detailed levels were relatively important and which seemed unimportant and of lower priority. In other words, we avoided committing ourselves to a concentration of effort at the lowest and most complex level in our hierarchy except when we felt that activity at that level might influence the goals of the project.

Related to this was our decision to halt the collection of data about individual behavior in any subpopulation of the project population when the data revealed little difference among members of the sample. We were able to do this because all data on individual behavior were collected by means of semi-structured face-to-face or telephone interviews. (For a variety of reasons, chief of which were the emotional fallout discharged by the project and the relatively small size of the project population, we chose not to use the self-report questionnaire as a means of data collection.)

Collecting Data

Our data-collection effort was guided by a very detailed outline of dimensions of the subsystems. This document was drawn up at the start of the project. At that time we had to assume we were naive and stupid and that this was not to be a steady-state system. So we made provisions for continually making additions to, modifications of, and deletions from, this document.

To give you a sense of this document, let me describe in some detail the dimensions of information we collected about each of the 64 task forces or teams. First, we were interested in the manner in which the team was formed, the key people involved in its formation, the manner in which the key people emerged, the professional roles of these

key people, their attitudes toward the formation of the task force and the Triple-T project, the processes in the formation of the team, and criteria employed in team formation. Second, we were interested in changes in team composition over time, the reasons for such changes, the patterns of changes, such as size and representation. Third, we wanted to know about tensions within the team, the objects of tension, particularly whether tension was related to people, to ideas, or to the project.

Once a design for a local project had been agreed upon by the team, we collected information on the actors in the design, on dimensions of the design such as themes, breadth, depth, complexity on the one hand and trainees, trainers, and training facilities on the other. We were also interested in impact, especially philosophy about impact, the strategy of impact, and directions of impact. A project design might also address itself to extra-educational matters, such as political and financial consequences of the training program.

Finally, there is the crucial matter of defining the respective responsibilities of the various parties involved in training. A general question was posed: How did the team fix the several separate but coordinated contributions of the university and the schools and the community to the training of the target groups?

Our attitude toward data sources was based on the principle of seeking information from individuals directly only when the information could not be obtained from other sources. Much of our information came from various documents. This project was rich in document production. Documents were issued by all components of the system for both internal as well as external purposes. For example, the National Advisory Committee devised a form for describing the program plans of a task force, to be filled in by a member of the committee when he made a site visit to the task force. The institutes issued periodic newsletters which described the plans and activities of the task forces they served. The institutes held periodic one- or two-day meetings of task force leaders at which presentations were made on project developments. Members of the evaluation staff were usually present at these conclaves, and their notes on discussions were added to the files as secondary documents.

A Problem of Semantics

There is no question that in doing this work we have been plagued by the term "evaluation." I thought I had made peace with myself in a distinction I had come to earlier, which is that we *appraise* individuals but *evaluate* programs. However, evaluation implies determining whether or not the Triple-T Project, the institutes, the National Advisory Committee, or the task forces were "good" or "worthwhile." We just could not construct suitable criteria for making such pronouncements.

Perhaps it would be more appropriate for me to use the term "assessment," for our task was not unlike that of psychological assessment in that we sought to grasp as adequately as possible the total pattern of information presented by the system under study. We may think of assessment as the development of a model or working-image of the system being studied (8). This model can contribute to, or be integrated with, decision making that affects the development of the system—that is, process control. In the Triple-T Project, however, the model or image was not intended for process control, but rather for process evaluation. That is to say, our assessment was intended to monitor events and activities so as to permit the *evaluation*, rather than the control, of the project.

Assessment of an exceedingly complex process or system is a fascinating and challenging task. It demands clarity of conceptualization of the system being studied. In addition, a high degree of flexibility, or adaptability, is needed by those charged with assessment because there is seldom an opportunity to replicate events or observations. We were confronted with uniquely changing patterns in the process by which the 64 task forces across the nation developed project proposals in interaction with one another and with other subsystems of the Triple-T Project over a period of several months. As I have already indicated, the plethora of minutiae that might have been monitored was both confusing and overwhelming. As a matter of fact, the unfolding of events was extremely fascinating and compelling, and we had to resist becoming transfixed viewers of the high drama.

One outline of the course of assessment, by Sundberg and Tyler (8), identifies four major stages: 1. the *preparation stage* during which the problem or system to be studied is identified and assessment methods are planned; 2. the *input stage*, during which information about the system, subsystems, and environment is identified and collected; 3. the

processing stage, during which the information collected is organized and interpreted; and 4. the *output stage*, during which findings and interpretations are communicated to others, and decisions are made about actions to be taken.

In the assessment of the Triple-T Project these are not discrete, sequential stages but may be thought of as streams of concurrent activities initiated with the project and ending with the publication of our final report. One stage or stream may have predominated at any one time and in the order given. But once initiated, each has continued as the assessment progressed.

Wiener (9) has noted that:

> to describe an organism we do not try to specify each molecule in it and catalogue it bit by bit, but rather to answer certain questions about it which reveal its pattern, a pattern which is more significant and less probable as the organism becomes, so to speak, more fully an organism.

The identity of each of the project subsystems is thus seen to lie not in its members, but rather in the continuity of pattern or process. The subsystems interact among each other within an extremely complex environment. Each subsystem is self-contained but interfaces with, or adapts to, the others. It is important to note that each may have, and did have, its own values and goals that are distinct from, and possibly incompatible with, those of the Triple-T Project.

If we attempt to apply cybernetic concepts to the highest level in the hierarchy of models in Figure 2—namely, the Triple-T Project in its entirety—the input from the environment consists of a "Program Plan" devised by a national planning group. This plan was communicated to the subsystems in different ways. There was a preponderance of oral communications and a multiplicity of written reports, only one of which is considered to be "official." Furthermore, though the communications were seldom incompatible, they were not always equivalent. It was no easy task to determine what information concerning the project was presented to, or perceived by, each subsystem. The nature of the project demanded that we monitor the communication of the Program Plan over its entire period. As a check on the accuracy of our information we decided to probe the subsystem periodically for their perceptions of the plan during and after the project-development period.

The output of the project was, of course, 64 project proposals and their evaluation by the National Advisory Committee. In a sense we had an evaluation within an evaluation. An obvious difficulty confronted us when we attempted to identify relevant information pertaining to the output of the total project. There had been no prior consensus on precisely what the output should be with regard to the content or aims of any proposal beyond the mandate given to the multi-sectored task forces. Each task force was to represent and bring together representatives of all sectors and identify appropriate needs, resources, and strategies to solve their problem. If we conceive of a task force as a cybernetic, or self-regulating, goal-seeking system which progresses toward its goal by utilizing information feedback concerning its activities, then we would predict the occurrence of considerable "searching" activities in the task forces. Self-regulation requires that the system's behavior be a function of both the feedback concerning its performance and a comparison of this performance with a criterion or goal. The stated goals of the total project—to the extent to which they could be identified—revealed few criteria of any usefulness for feedback to the task force itself, the institute, or the National Advisory Committee.

Unfortunately, this placed the team in a problem-finding and problem-solving—or goal-identification as well as goal-seeking—situation. This in itself may not have been a handicap to the majority of task forces, but it may have reduced the catalytic effect of the institutes.

An additional problem arose for the task force when it sought to solve the puzzle of what the National Advisory Committee would look for in project proposals when it evaluated them. This was also a puzzle for the national committee. When it tried to develop criteria it became embroiled in the very matters that task forces were supposed to be debating. This was because the composition of the national committee was also multi-sectored. We watched with fascination as the national committee became the sixty-fifth task force for three exciting days.

If we restricted our notion of output to the preparation of 64 project proposals, we could simply count documents or summarize them. If, on the other hand, we were to consider the project proposals accepted without reservation by the national committee and recommended for submission to the Office of Education, as well as other decision categories of the committee evaluation, we would risk

assuming the validity of their evaluation procedures in the absence of any criterion external to those of the committee.

The result of our decision concerning the delineation and assessment of the output of the project consists of the presentation of an abstract of each of the 64 project proposals and a content analysis of those proposals that were approved without restriction and a sample of proposals rejected outright by the committee. This represents a total of 40 percent of the proposals submitted. This procedure allows us to compare the extreme category groups of proposals and make inferences as to the criteria implicitly identified but not necessarily used in the national committee evaluation.

The ideas of the project that are contained within the structure, especially those of parity and verticality, are the subject of special studies. These studies are being conducted at both the logico-philosophical and empirical levels of analysis. The evaluation group was organized to include people with experience in a variety of domains so that nonempirical studies could be done. Just as we didn't conceive of our work as being psychometric-centered, we also didn't view it as particularly empirical-centered. We defined our responsibility as broadly as possible. Thus, when studying the idea of parity we examined the assumptions underlying this idea, we looked at the idea in the light of relevant social science theory and knowledge, and we studied what happened to the idea of parity from the time the national planning group developed criteria for forming task forces until the project proposals were sent to the Office of Education some seven months later. Empirically, we dealt with the idea of parity in four senses: 1 as an idea—that is, how it was received and discussed; 2. in terms of composition of task forces over time; 3. in terms of interactions within task forces; 4. in terms of the proposed projects.

Other properties of the project design are also topics of special study. The educational functions of the institute, the management of the project by the National Advisory Committee, and the alteration of methods of arriving at proposals are examples of such properties. With respect to the latter, it did not come as a complete surprise to realize that much of the early noise and tension in the system was traceable to the new way in which the funding game was being played. I spec[illegible]efer here to much less structure being imposed from witho[illegible] who had become conditioned to a much greater degree of stru[illegible]ound themselves with ambivalent feelings about the open-ended structure of this project and without the skills required by it.

Finally, a few words about values and objectivity in evaluation. Although Figure 2 doesn't show it, we consider the evaluation group to be a system external to, but interacting with, all subsystems of the project. We did not see an inevitable conflict between our sentiments toward the project and our firm resolve to be as objective as possible. Nor were we concerned that cognitive and affective changes in us, which were traceable to our project experiences and which, I might add, were very pronounced and significant changes in a number of us, should cause us to declare ourselves no longer capable of being rational and objective. We succeeded early in communicating the role and purposes of evaluation in this project and thereby avoided creating tension among those who think of evaluation as a threat. That we experienced no difficulty in obtaining documents, conducting interviews, and tape-recording conferences seems to me to have been proof that we were well-regarded and to be trusted. That our advice and opinion was sought on many occasions was also inevitable. We had no trouble sensing the criterion that would govern our response. Where explanation and clarification of Triple-T Project properties and objectives was being sought, we gave freely of our time and energy. Where observations on local project designs were sought, we had no trouble declaring these requests to be out of bounds.

The funding of our work by the Consortium of Professional Associations for Study of Special Teacher Improvement Program (CONPASS), helped give us a respected identity of being neutral but wishing the project well.

Conclusion

In this paper we have tried to give you a sense of the origins and scope of a national project and a sense of some of the more salient issues facing an evaluation group that is monitoring and assessing this project. Without saying it directly, we meant to suggest that our task was a difficult one, that a project of this scope is quite formless and dirty and certainly ever-changing, and that our methodology is far from elegant. Neither the project nor our work could be neatly packaged. Whether the field of educational program evaluation will ever be liberated from its primitive basement and garage workshops will depend on our ability to generalize from the ever-increasing number of opportunities we have to evaluate these large-scale efforts.

REFERENCES

1. Beer, S. *Management science: the business use of operations research.* Garden City, N.Y.: Doubleday & Co., Inc., 1967.

2. Bigelow, D. N. The gatekeepers. An audio-visual presentation to participants in the TTT Institutes, January 22—February 29, 1968.

3. Bigelow, D. N. An overview of the Division of Educational Personnel Training Programs for Teachers of Teachers. In *Teacher Education—Issues & Innovations.* Washington: American Association of Colleges For Teacher Education, 1968.

4. Haubrich, V. *But . . . which way do we go?* Madison, Wisconsin: Dept. of Educational Policy Studies, University of Wisconsin. Unpublished manuscript, 1967, 6.

5. Olson, P. A. The tri-university project in elementary education. In *Teacher Education—Issues & Innovations.* Washington: American Association of Colleges For Teacher Education, 1968. P. 67.

6. Smith, B. O. On the preparation of the teacher of teachers. Paper presented to the Midwest TTT Institute, Michigan State University, East Lansing, Michigan, February 26, 1968.

7. Smith, B. O. Some basic issues in the preparation of teachers of teachers. In *Teacher Education—Issues & Innovations.* Washington: American Association of Colleges For Teacher Education, 1968. P. 74.

8. Sundberg, N. D. and Tyler, Leona E. *Clinical psychology: an introduction to research and practice.* N.Y.: Appleton-Century-Crofts, 1962.

9. Wiener, N. *The human use of human beings: cybernetics and society.* Garden City, N.Y.: Doubleday & Co., Inc., 1954.

Cost-benefit Analysis and the Evaluation of Educational Systems

J. ALAN THOMAS
University of Chicago

Introduction

Throughout the world, the demand for education is increasing more rapidly than the supply of classrooms and teachers (7). Furthermore, both at home and abroad, traditional notions of the manner in which education should be distributed are being questioned.* These pressures for more education and for a more equitable distribution of knowledge are creating pressures on our educational systems. The efficiency of educational systems at the national, state, and local levels is, therefore, a matter of prime concern to those who are interested in education's contribution to a viable society.

If educational systems are to be improved, procedures must be developed for evaluating them. Evaluation can then provide guidelines for the improvement of efficiency.

Psychologists who specialize in testing and evaluation have made major contributions to the development of a scientific base for the improvement of education. Breakthroughs which may be expected in the application of their findings toward the improvement of policy making will, we hope, have a strong impact on the efficiency of educational systems. Psychometrics does not of itself, however, provide a sufficient basis for *school system evaluation*. Consider, for example, the task facing a school survey team charged with the evaluation of a large educational system. The team will wish, as part of its data, to have compilations made of achievement and intelligence test scores

*See, for example, Arthur E. Wise (14).

in the school system. However, these scores will probably reveal more about the social and economic characteristics of the community than about the effectiveness of the educational system (6).* Nor do differences in test scores among schools within an educational system provide (without the use of additional information) sufficient criteria for school system decisions such as those leading to the manner in which resources shall be allocated.

This paper is based on the assumption that educational organizations must be considered as "open" systems for the purpose of evaluation. Such systems are considered to be in a constant state of interaction with their environment. They absorb energy (in the form of various inputs—including incoming students) and contribute energy (outputs, including graduates with developed productive capabilities) (8). Educational and other types of systems may therefore be evaluated by comparing their outputs with their inputs. This comparison of contributions with costs constitutes an evaluation of school systems' social productivity.

Evaluation by Economists

Some of the most successful attempts to evaluate educational systems have been made by economists. There have been several reasons for their success. In the first place, the professional interests of economists lead them to look upon educational organizations as open systems, which are linked to the total economy through a set of inputs and outputs. In the second place, their concern for costs and benefits provides a context within which evaluation can take place. In the third place, economists have developed analytic procedures which enable them to compare costs and benefits, even though both are incurred over relatively long time periods.

The economist's "evaluation" of an educational system is so different in appearance from the evaluation of measurement experts as to warrant some further explication. Table 1 shows the rates of return calculated for high school graduates, college graduates, and corporate manufacturing firms in comparable years.

This analysis uses measures of rates of return that are similar in

*For a critical evaluation of the Coleman Study, see Samuel S. Bowles and Henry M. Levin. (3).

meaning and in derivation in the measurement of the efficiency of both educational and manufacturing systems. These comparative data suggest that high expenditures for education, especially at the high school level, would be desirable from a total societal point of view, since investment in human capital seems to have a higher rate of return than investment in physical capital (12).

A similar conclusion is reached by Denison, whose research suggested that 23 percent of the economic growth in the United States between 1929 and 1957 was due to the effects of education (10). These results (like other research findings) are affected by the assumptions they incorporate. The assumptions must be examined critically, and the studies need to be replicated as new data are available. However, these studies suggest that the economist's tools for evaluating educational systems cannot be ignored by educators. Furthermore, the *results* of their research have important implications for social policy.

There will be some immediate objections to permitting economists to share the task of evaluating school systems. It is often thought that economists are interested only in educational benefits that can be expressed in terms of dollars and cents. This is unfair; economists are willing to regard education as consumption as well as investment, and accept both monetary and nonmonetary benefits as being important. However, it is true that their analytic procedures work best when the variables can be expressed in monetary terms, and there may in leed

Table 1*

Sector	*Rate of Return*	*Year*
High School Graduates: White Males after Personal Taxes	28%	1958
College Graduates: White Males after Personal Taxes	14.8%	1958
Corporate Manufacturing Firms: After Profit Taxes but before Personal Taxes	7.0%	1947–57

*Source: Theodore Schultz (12)

be a bias in favor of studying the material benefits derived from schooling.

Up to this point, their greatest success has been in studying the costs and benefits associated with the national educational system. Although these findings are important, they do not provide the information which is necessary in improving decision making within local school systems, or even in state departments of education. In order that cost-benefit analysis of local school systems may be conducted, some merging of the knowledge, skills, and interests of the economist and the measurement expert in education may be necessary. Before touching upon some aspects of this task, I should like to discuss briefly the techniques of cost-benefit analysis.

Cost-benefit Analysis in Education

Economists have made an important contribution to the study of school systems by their development of the human capital concept (1). Human capital (analogous to physical capital) refers to the developed productive capabilities of human beings. Like physical capital, human capital is a produced good, formed as a result of formal and informal schooling—in the home, in school, and in other organizations. The formation of human capital requires the use of resources—whether they constitute the time of a mother who foregoes other activities, the purchased time of teachers, or the efforts of a factory supervisor. The benefits associated with human capital typically constitute a stream of financial and other rewards an individual may receive over the remainder of his lifetime.

Not all benefits and costs associated with education can be expressed in terms of dollars and cents. However, monetary costs and benefits are important. Increased education results (on the average) in additional income. It also results in monetary costs—in the form of out-of-pocket payments and also a loss of earning power during the period while the student attends college or senior high school.

One of two main procedures is usually used in cost-benefit analysis (1, 4). One is to reduce the stream of benefits and the stream of costs to a *present value* at a given year by using the mathematics of compound interest and compound discount. The other is to calculate a *rate of return* that would equate the stream of costs to the stream of added income. An investment is defined as "worthwhile" if the present

value of the benefit stream exceeds the present value of the cost stream (at a selected rate of interest) or if the rate of return exceeds some externally determined figure.

Economists take nonmonetary as well as monetary costs and benefits into consideration in their studies of human capital. In other words, education is considered as *consumption* as well as *investment*. Hence, the stream of benefits will include the many satisfactions that a good education provides, while the stream of costs includes foregone leisure as well as the foregone earnings associated with schooling. These nonmonetary costs and benefits are, of course, more difficult to deal with than those that can be expressed in monetary terms.

Not all costs and benefits are directly attributable to the individual who is the direct recipient of educational services. Some are *external*—that is to say, they are attributable to people other than those immediately concerned, namely, the students and their families. For example, the entire society benefits from the higher rate of economic growth said to result from education; employers benefit from having a pool of educated potential employees; scholars benefit from living in a society where there is a wide variety of books and magazines; music lovers benefit from a society in which many other people have also been educated to produce and consume music. Similarly, people other than the student help to pay the cost of education (13). *Social* costs or benefits are defined as the sum of the private effects and the external effects.

Rigorous cost-benefit studies of local school systems are still far from practical, for the following reasons:

1. Historical information about the postschool income streams of graduates of a given school system are extremely hard to obtain. Furthermore, income streams of past graduates are not completely satisfactory for evaluating decisions made today, under different circumstances.

2. Large-scale migration in and out of our major school systems hampers this type of analysis.

3. This steady-state analysis is partially, but not altogether, irrelevant to the decision-making dynamics of school systems.

Despite these limitations, the cost-benefit model provides a guide in the determination of the kinds of data that should be gathered.

Benefit and cost data should be obtained and should include the following:

1. Information about the post-secondary school academic and occupational careers of students: Percentages of students going on to four-year and two-year colleges, or to postsecondary vocational institutions; proportions who after or before graduation become employed in different categories of occupations; income subsequent to leaving school; unemployment of school leavers—these data are all relevant in estimating the effectiveness of an educational system. In large school systems, these data should be disaggregated, according to secondary school and to home background (including race, occupation of fathers, economic status of parents). This provides an indication of the school system's effectiveness in dealing with different subgroups of students.

2. Information about costs: The historical cost pattern provides a clue in the determination of productivity trends. If costs have been rising while outputs have fallen or remained constant, productivity also has decreased—unless other factors such as a changed school population have intervened. The costs of educating subpopulations should be calculated as one measure of the allocatory procedures that are at work. For improving decision making, it is also necessary to have detailed information about the unit costs associated with various school programs, changing prices of inputs, and (as one indication of the rationality of the system) the cost of obtaining and processing information. The purpose of cost accounting is not merely to ensure economy. Costs are part of the basic structure of decision making. The cost of providing a given curriculum or of implementing a desired teaching methodology must be gauged in part in terms of the other curricula or the other methodologies that must be foregone, since total resources, at any given time, are limited.

For some purposes, the cost-benefit terminology is too restrictive, since it implies an emphasis on monetary variables with positive and negative signs. The exchange between an open system and its environment can better be expressed in terms of inputs and outputs. Inputs include all the contributions of the environment to the system, including some that cannot be described as costs. Outputs include the

products of the system. The following are some kinds of input and output information that are relevant to school system evaluation:

Outputs

Post-high school experience of graduates

Years schooling completed by students (includes a study of dropout and retention rates)

Achievement in the various subjects at the various grade levels (Ideally, achievement should include measures of affective as well as cognitive learning)

Rates of promotion and nonpromotion

These data should be disaggregated by school and by characteristics of the student body, including race and socioeconomic standing.

Inputs

Students: (Including their home background. Measures of achievement press in the home of the type developed by Dave and Wolff [9] would be desirable, as well as the usual social and economic data.)

Teachers: Their background, training, experience. Attitudinal measures seem essential, in view of a possible relationship between teacher expectation and student achievement.

Administrators: (similar to teachers)

Physical inputs: School buildings, equipment (especially technological books, and so on).

Management: (Here we include the seeking and utilization of information in decision making, the utilization of information from outside the system, and the use of the more sophisticated decision-making management procedures now being developed.)

Again, data must be disaggregated. For example, there should be detailed information about the distribution of teachers among types of school. This should include attitudinal data, such as differences in attitudes of teachers in all-Negro and in white schools.

A careful analysis of this information can lead to some conclusions about school system productivity. Historical trends are, of course, important. A study of input and output changes within the system may point to hypotheses concerning possible strategies for educational

improvement. One superficial conclusion might be that an unequal allocation of funds among the schools of a school system (with, for example, less money being spent on the education of Negro children than of white children) is unproductive, in terms of the total social cost-benefit relationship. A more sophisticated analysis, including a study of the teaching body assigned to the various schools of the city, might suggest that unequal opportunity has deeper roots than the way in which money is allocated.

The Evaluation of Proposed Changes

The careful analysis of an organization's inputs and outputs is similar to the development of a detailed profit-and-loss statement of a business firm. However, inputs and outputs of educational systems usually cannot be stated in the same units, and hence the two sides of the equation cannot be compared mathematically. Judgments about efficiency can be made, usually in the form of hypotheses about the effects that might be associated with changes in present procedures.

Alternatively, cross-sectional studies among a number of school systems may provide information about the marginal effects of the various input variables. However, cross-sectional studies do not, as is well known, indicate causation. The logical leap from the marginal relationships implied by multiple-regression analysis to the improvement of efficiency in practice calls for judgments to be made about the effect of proposed changes in a particular situation. Hence, the analysis of changes over time in inputs and outputs in a given school district probably provides the most useful way in which an organization may obtain the data basis needed to improve its productivity.

This type of evaluation is based on the examination of anticipated productivity increases associated with change, and is therefore dynamic rather than static. These changes will include some reallocation of money among the various inputs used by the system, the possible redeployment of inputs, and in some cases the use of new inputs.

This analysis is governed by a well-established rule of economics. This is that productivity will be maximized when money is distributed among the purchase of inputs in such a way that the marginal product of each of the inputs is the same. Thus, efforts should be made to determine whether any possible reallocations among inputs (such as

spending more money for books and less for floor wax) will improve productivity.

In order to illustrate this procedure, this paper discusses three proposals made in a recent school survey (11). The purpose of this analysis is not to justify these proposals but to illustrate one use of input-output analysis.

1. One proposal, involving a suggested major allocation of resources, was that increased emphasis be placed on preschool and early childhood education. Preschool programs for a large section of the preschool population were proposed. These would be associated with a restructuring of kindergarten and primary education. New inputs would be needed, in terms of additional teachers, classrooms, materials, and books for young children. In addition, existing inputs would be improved through extensive in-service education for existing teachers.*

From the point of view of the system as a whole, this would mean spending a larger proportion of the total school budget for early childhood education. It would imply emphasizing one type of input (teachers of young children) at the expense of another type of input (teachers of secondary children). Of course, these are marginal changes; it was not proposed that there be fewer secondary school teachers or that their salaries or training levels be changed, merely that there be a new emphasis.

On the benefit side, the hypothesized effect is that this reallocation of resources would lead to an improved total output over a period of time. It is hypothesized that if this program were to lead to the firmer acquisition of basic skills by young children, they would learn more in high school, require less remedial attention, stay in school longer, and earn more and learn more after leaving school.

To be sure, the long-term results of such a change could not be anticipated at the time of its implementation. A careful use of feedback would be needed, in order that gradual year-by-year improvements be noted, and so that the strong aspects of the proposed changes could be emphasized and the weak aspects eliminated.

2. A second proposal, with important cost-benefit implications, was

*The empirical basis for the proposal is Benjamin S. Bloom's *Stability and Change in Human Characteristics* (2).

that attention should be given to the restructuring of teachers' roles, with emphasis on greater role specialization. The proposal would include the hiring of teacher aides and other para-professionals. It would also include provisions for creating new roles, such as those of team leader and master teacher. This suggests a major reallocation among inputs, with the possibility of paying higher salaries to the best qualified teachers. It suggests improving efficiency through assigning to individuals the types of responsibility for which they are most qualified. Clearly, in terms of our previous definition, this attempt to pay teachers more nearly on the basis of their marginal contribution is one potential way to maximize total system productivity. However, changes in salary practice would have to be accompanied by ongoing studies of changes in costs and benefits.

3. A third recommendation was to obtain a different type of input, in the form of staff personnel at the central-office level. School districts tend to be staffed mainly by line personnel, whose task is to give and to carry out orders. Furthermore, line personnel tend to have a common training—usually through their experience as teachers and subsequently as administrators. Staff people would have as their duties: (a) obtaining and analyzing data, (b) long-term planning, and (c) evaluating. They would be chosen from a pool other than traditional administrators, and might include social scientists or other people with noneducation backgrounds.

One purpose of this recommendation was to reallocate resources in the direction of providing more and better knowledge at the point where decisions are made. To be sure, additional knowledge does not ensure better results; it is probably a necessary but not a sufficient condition for improvement. However, if accompanied by proper monitoring and feedback provisions, better knowledge should result in improved performance.

Summary

This paper has described an approach to school system analysis based on a study of costs and benefits. The analytic procedures suggested in the paper are quite general, since the state of the art is not yet such

as to provide a well developed methodology or even a tight, consistent theory.

Several important aspects of the topic have been given short treatment, as being peripheral to this paper. For example, the study of unit costs is obviously important, but has been only touched upon in this paper. Operations research techniques, designed to generate the knowledge needed for the improvement of cost-benefit analysis, have not been discussed. This is an important area in which much more work needs to be done if schools are to provide the services which society demands with the money which will be available.

REFERENCES

1. Becker, Gary S. *Human capital.* New York: Columbia University Press, 1964.

2. Bloom, Benjamin. *Stability and change in human characteristics.* New York: John Wiley & Sons, 1962.

3. Bowles, Samuel S. and Levin, Henry M. More on multicollinearity and the effectiveness of schools. *Journal of Human Resources*, 1968, 3, 3, 393-400.

4. Bowman, Mary Jean. Schultz, Denison, and the contribution of "eds" to national economic growth. *Journal of Political Economy*, 1964, 72, 5, 450-464.

5. Bowman, Mary Jean. The human investment revolution in economic thought. *Sociology of Education*, 1966, 39, 2, 111-137.

6. Coleman, James S. Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office, 1966. Burkhead, Jesse. *Inputs and outputs in large city school systems.* Syracuse: Syracuse University Press, 1967. Thomas, J. Alan. Efficiency in education: the relationship between inputs and outputs in a sample of high schools. Unpublished Ph.D. dissertation. Stanford University, 1962.

7. Coombs, Philip H. *The world educational crisis.* New York: Oxford University Press, 1968.

8. Daly, Herbert E. On economics as a life science. *Journal of Political Economy*, 1968, 76, 3, 392-406.

9. Dave, Ravindrakumar H. The identification and measurement of environmental process variables that are related to educational achievement. Unpublished Ph.D. dissertation, Department of Education, The University of Chicago, 1963. Also Wolf, Richard M. The identification and measurement of environmental process variables related to intelligence. Unpublished Ph.D. dissertation. Department of Education, The University of Chicago, 1964.

10. Denison, Edward F. *The source of economic growth in the United States and the alternatives before us*. New York: Committee for Economic Development, 1962.

11. Midwest Administration Center *Cincinnati school survey, volume 1*. Chicago: Midwest Administration Center, The University of Chicago, 1968. Pp. 47-48.

12. Schultz, Theodore W. Resources for higher education: an economist's view. *The Jour..al of Political Economy*, 1968, 76, 3, 337.

13. Weisbrod, Burton. External benefits of public education. Princeton, N.J.: Industrial Relations Section, Princeton University, 1964.

ise, Arthur E. *Rich schools, poor schools*. Chicago: University of Chicago Press, 1968.

Luncheon Address

A Customer Counsels the Testers

SIDNEY P. MARLAND, JR.
Institute for Educational Development

The institution of testing has become an Establishment. As such, it has become a victim of its own success. Even as the Vatican. And the Pope. The Supreme Court, boards of education, universities, school administrators, the Democratic Party—these are other establishments.

To be an Establishment nowadays is to invite the wrath of a fair portion of our society on almost any issue. Dr. John Gardner has reflected on this matter of the Establishment. In an address he gave at Cornell University this past June (I hesitate to paraphrase Gardner with his Churchillian prose), Dr. Gardner said that if an observer could take a pill that would thrust him ahead 300 years from now and permit him to look back upon the history of the latter half of the twentieth century in America, he would find that this was a time when all of our established institutions came under great challenge because of the unloving critics of institutions and the uncritical lovers within the institutions. The critics not responsible for their institutions began to tear them down, and those who loved their institutions so dearly as to be uncritical of them did indeed allow them to languish. As this pill continued to work, it was observed that the wise people in the latter part of the twentieth century began to change: The unloving critics became somewhat loving and concerned, and those who were the uncritical lovers of their establishments became critical and constructively set about corrective action.

Today, I would say that the testers are becoming their own loving critics, and this is as it should be. For example, the College Entrance Examination Board will soom publish a very candid and thorough assessment of all the current criticisms of testing. Here is a passage from the report:

1968 Invitational Conference on Testing Problems

Standardized tests have been a source of considerable controversy over recent years. Growing competition for jobs, for admission to college and for educational opportunities in general has led to an intensified search for better ways of evaluating capabilities and aptitudes for identifying intellectual potential at earlier ages.

This great reliance on standardized tests has led a number of scholars and others to raise important questions about the validity of tests being used, about their effects on those who take them and on the society that uses them to differentiate among human beings . . . .

Three years ago, the College Board established a major commission to investigate itself—if you will, to be a critical lover of this Establishment. Richard Pearson, President of the College Board, in initiating the commission, said that the commission members "should undertake a thoroughgoing appraisal of existing tests . . . and in light of the future needs of admissions." This independent twenty-one member commission is now hard at work, having drawn upon the counsel of the most earnest critics in pursuit of its broad mandate from the College Board.

In my approach to the task that has been put to me today, I should like to make it clear that I do not pretend to have competence in the field of testing and all of its many ramifications. I do hope to present the useful observations of a school administrator who has been a customer of the test makers for a long time, who has had to defend budgets calling for hundreds and hundreds of thousands of dollars' worth of tests as well as salaries of staffs to administer those tests, and psychologists and counselors to interpret them and use them wisely with young people and parents. It has also been my task to persuade boards of education as to the usefulness, importance, and the need for a testing program in school systems. Thus, as a consumer of test products, I am in a position to offer some insight as to those criticisms that now surround the Establishment of Testing. In this paper, I will cite some of these criticisms and respond to them as a person claiming competence only as an administrator and a consumer.

Excessive Testing

A criticism that I would offer, and one that is on the lips of many of our critics, is redundancy—the multiplicity and excessive amount of testing that goes on in our schools. I think there is a great deal more test-

ing going on now than there was perhaps even five or six years ago, but I am not sure that there is any evidence to show that it is all happening to the same child! I think that institutions, schools, industries, government, and other consumers of tests are testing more. I think the schools—broadly speaking—are consuming more tests, and more testing and measurement is going on. I think this is probably a product of government intervention, the encouragement of the National Defense Education Act, and increasing respect for some of the applications of testing. At any rate, testing is reaching young people it has not reached before.

I feel that there is not an increase in testing in a given school that has been testing for some time; there may indeed be some decrease in testing as teachers and administrators and counselors become more sophisticated in using test information. I think, for example, that I could predict with some confidence that over the next two or three years the duplication between the National Merit Scholarship Testing Program and the Preliminary Scholastic Aptitude Test will be eliminated. Since it is clear that a very high correlation exists between these tests, I believe that they very likely will soon be one and the same, and that educators will agree that they can get the same kind of information from one test instead of two. As we become more certain of what tests can and cannot do, there will be an increased efficiency in their uses. I am speaking as a school administrator who has to justify these things in his budget and justify teacher time and pupil time.

Another point that is noteworthy among the current criticisms of testing is that the test users—namely, those of us in the schools—have extended the function or the implied function of tests beyond the intent of the test makers. This should not be a criticism of the test makers. David Goslin has done some very useful investigating into this whole business of the changing world of testing. According to Goslin, a great deal of the criticism surrounding tests relates to those concerned with college entrance selection and prediction, and yet, "considering only school testing, a very strong case can be made for the fact that standardized tests given in elementary schools have a potentially greater impact on both pupils and schools than do college admission tests.

"The tests are used for many purposes (by the schools), even though they may be primarily called the College Entrance Examination Board's Scholastic Aptitude Test, scholastic aptitude being all that the test makers intended this test to be."

Goslin discovered that principals acknowledge using these tests for a great many other things, such as the following:

1. the grouping of students according to some kind of homogeneity in classification within schools;
2. a basis for determining pupil strength and weaknesses for remedial purposes;
3. a basis for providing the pupil with information about his abilities as a guidance function;
4. a basis for evaluating effectiveness of teachers (unfortunate, but in many cases probably true);
5. a basis for evaluating the appropriateness of the curriculum as it relates to overall school system plans and effectiveness.

Some of these are probably not appropriate uses for test materials. Moreover, if the test is being used for other purposes, it should be known and should not be charged against the test makers. These uses should be optional functions within a given school system.

Another criticism is the one that says standardized ability tests measure only a few of the characteristics of the human beings under instruction; creativity, social responsibility, motivation, physical effectiveness, and many other characteristics are untouched. To give concentrated attention to the early intellectual or academic measures, these critics charge, is to distort the evaluation of any human being.

I happen to feel in full accord with those who make this observation, but I know that the test makers themselves are very much aware of this problem, and I know that they are going about the business of trying to solve it.

One of the dilemmas facing the test makers, I am sure, is that as they try to expand the test to take on qualities beyond the academic or intellectual, they have to deal more and more with some order of subjectivity in a value system of some kind. The paradox is that if we are searching for ways to provide equality of opportunity, especially for our services to the deprived child, the more objective our measures are, presumably the more honest our appraisal. This puts us in the dilemma of relying perhaps more and more on measures of finite things such as measurable achievement and measurable intelligence as distinct from the more subjective values such as attitudes and social responsibility and creativity and the things that make up this human

being, quite apart from academic learning.

In our search for sound and thoughtful criticism of present testing, the Commission on Tests of the College Entrance Examination Board has sought the counsel of many outspoken scholars on the subject. Kenneth Clark was among those who contributed to their deliberations on this concern for a wider spectrum of human qualities than our present pattern allows. He calls for a new approach to testing—one that will allow for experimenting with the testing of social values. "These are legitimate components of the education process," says Clark.

The Question of Testing Intelligence

Intelligence testing, I am sure, is one of the great concerns of many scholars. Alexander Westman, writing in *American Psychologist* last April, said "There appears to be no more general agreement as to the nature of intelligence, or a more valid means of measuring intelligence, than was the case fifty years ago. Concepts of intelligence and the definitions constructed to enunciate these concepts are found by the dozens—if not, indeed, by the hundreds—with so many diverse definitions of intelligence that it is perhaps not surprising that we cannot agree on how to measure intelligence."

Formalization and overemphasis on a numerical or quantitative score is another criticism. Test makers and test users continue to wrestle with this question. If the IQ score is, to those in the field of testing and measurement, a rather unreliable figure (and I think you will say it is), then let us stop using it and let scholars and social scientists, psychologists, psychometricians, and others deal with this in their own privy councils and research. But let *us* stop attaching dubious exactness to an unreliable measure that sounds very reliable to people like teachers and parents and children and school superintendents.

The users of the tests, the translators of the test, the teacher, the parent, are given to feel that the IQ is something exact when no one claims it to be exact. Therefore, have we thought of setting it wholly aside and perhaps confining ourselves to the use of profiles or stanines in the placing of the child for a more easily translated and more comprehensible description of his characteristics and his needs? Quoting again from Westman in the *American Psychologist:*

All ability tests—intelligence, aptitude and achievement—measure what the individual has learned, and they often measure with similar content and similar process.

Doesn't this again suggest there could be some kind of a composite measure surrounding the child—again with the increased flexibility that we derived from data processing—so that we could set aside the apparent exactness of finite figures in describing human beings and maybe come up with a profile that could, perhaps with a stanine format, provide some sort of a card-punch expression of each child's characteristics?

It may all be in one category—intelligence, aptitude, and performance—if, as Westman says, they are so much alike and they merely measure what the human being has learned as a result of systematic and nonsystematic experiences.

It would also be useful if on this same stanine there were modes or norms, norms of the universe, and perhaps norms for the kind of subcommunity in which the child lives, the kind of people with whom he relates, and the kind of environment in which he finds himself. It is desirable that he and the teachers have some idea of how he relates broadly to the immediate and the larger population, without trying to impose an exaggerated exactness in which none of us has confidence, especially those of you engaged in constructing the measuring device.

I would underscore this topic as an important part of this paper. I would emphasize the need for inventing a new way of providing a profile of the child's ability and performance against the environment in which he is living and working. This new device should be so constructed that it can be easily communicated to counselors and teachers, allowing for varying degrees of expertise. It must be so designed that it can be fully comprehended by them and, in turn, readily translated to parents and children.

Moving to another criticism of testing, there is the charge that testing is a mechanism for the self-fulfilling prophecy—that a test score earned at an early age or, indeed, at any age, is such a monumental piece of evidence that it cannot be overcome either by the child or teacher.

I am not so sure that this criticism is valid. I don't think that our teachers are that poor. I don't think that teachers use the test with such vehemence that, indeed, it becomes a self-fulfilling prophecy. I am aware of some of the experiments that have shown teachers to respond

in this way, but I would say this is a matter of in-service training of the teacher rather than a fault of the test. It is a matter of administrative management in the school system to determine how tests should be used and how their meaning should be translated to positive and constructive uses. Teachers need much help in using test results, and school leaders should be responsible for insuring such help.

Most of us are familiar with criticism of multiple-choice tests. The charge is usually that the child is dehumanized by standardized testing, that we remove from him the opportunities for creativity, for critical thinking, and for imaginative response to questions by limiting him to the little marks within the constraints of multiple choice. This is probably a good scholarly criticism from the viewpoint of various academic disciplines. But I would say again that the critic may not know what else is going on in the schoolhouse.

The use of standardized tests is a small part of the teaching system. Teachers search constantly for opportunities to release creativity in children, to provide challenging teaching and learning environments in which children can practice critical thinking. In any reasonable schoolroom there are a great many other opportunities for a child to be treated as a lively and distinctive human being, apart from the infrequent constraints of the multiple-choice experience.

Do Tests Test the School?

There is a constant anxiety and concern that tests also test the school system and the teacher. This is probably true to some degree. We evaluate ourselves; we match ourselves, individually or corporately, with some kind of norm and say "We didn't come off so well" or "Aren't we good?" Many teachers—especially young teachers and sometimes not-so-young teachers—fear the results when their children are tested.

We know the unhappy instances of fraudulence in testing in one form or another, reflecting the anxiety of teachers. Grooming for tests, manipulating results, and even making false reports are occasional products of this unhealthy anxiety. This is completely wrong, but it is not a fault of the test itself. It is a fault of the system of teacher management. This is a dilemma that the test makers may wish to consider as a new challenge. We do indeed have a need in education for assessing the effectiveness of teachers. The pupil test is a shabby

and unworthy alternative to the assessment of teachers. But we have never faced up to systematic and objective teacher evaluation. I suppose we don't have any valid way of doing it.

Almost any scholar of testing and measurement probably would agree that this is a dangerous thing even to consider, but I would ask that the test makers think of some ways to test the effectiveness of teaching. This is part of education's job. We are trying to increase greatly the effectiveness of teachers. We must discover ways for increasing the productivity of teachers. We must provide for teachers increased means to pour out their professional talents and power, and we need some valid way of calculating benchmarks in this process. This is not for reasons of ranking or dismissing or rewarding or punishing; these are not appropriate purposes for measurement. What we need is a set of measures that can identify superior teaching so that its characteristics can be identified and replicated in others. Until we have some design for measuring effective teaching other than the feelings of a distant and occasional supervisor who says she is a great teacher, or a group of parents who say she is splendid or awful, we don't have very much. I would ask you to ponder this subject as a very real need in education at a time when accountability is more and more a mandate upon the schools.

This raises another question, which concerns the differences in types of groups tested. The teacher in a ghetto may be doing a far better job of teaching if you have some god-like way of assessing what that teacher is doing. The class may be performing two years below normal on the average, with some as many as four years below norm, say, at fifth grade, yet you and I, as experienced observers, might say the teacher is doing a superior job of teaching. In some other community, with a very favored population, there could be a pitiful job of teaching going on yet with children performing two or three grades above norm.

Getting back to my point about the different kinds of communities, I wish that we could have some better measurement tools from those people who analyze tests. We need help comparing like things when we deal with norms. For example, let us say in Pittsburgh we are dealing with a ghetto-school community. I should like to know how those children in that school community, a subcommunity within a large city, compare with the same kinds of children in the same kinds of subcommunities in Chicago, Milwaukee, Cleveland, and Cincinnati. And I would want the same unit of comparative data for favored

communities as well as for the total population!

Some tests, particularly the personality test, are considered invasions of privacy. Perhaps they are. I think we are in a time when we are going to have to accept the invasion of privacy. I think that when, for good reasons, all of us at one time or another get around to having our fingerprints taken and recorded somewhere, that is invasion of privacy. And I think that when the credit rating of 120 million Americans is programmed into a computer with an access code available to 27,000 different establishments that can learn about our buying habits and our paying habits—that, too, is an invasion of privacy. But I don't hear anybody rising up and shouting about that.

We might as well decide to live with conditions of data gathering and data retrieval that are presumed to be of sufficient value to offset the discomfort of being "invaded." As a school administrator and parent, I am ready to accept the advancements of science as a resource to the schools provided useful ends are served by the data scientifically gathered.

A critical problem for school administrators has to do with ways in which test results are reported. We are going to go through a period of considerable debate in our country over the publication of test results by schools or by individuals or by neighborhoods. I think it is a wasteful, undesirable, and unethical practice to disseminate test results by schools or neighborhoods. I don't think it accomplishes a thing. But we are going to go through that debate, and we ought to have some ground rules for guiding school policy makers in this difficult field.

There are those who fear the control of the curriculum through testing. A study recently made by Roald Campbell of Chicago with Roderick McPhee has shown that, indeed, those national symbols of power, prestige, and authority such as the College Entrance Examination Board, the National Merit Scholarship Corporation, and the National Science Foundation have indeed influenced curriculum. I do not think this is necessarily bad. I think that we are living in a time when there have to be some central standards, expectations, and goals that give uniformity and consistency to all the things we are doing in the schools without necessarily being dominated by these external standards. But I think that it is unfortunate that we have to leave everything to be determined at the local level.

I think it is doubtful that the expectations and mandates of our people can be met in a reasonable period of time if we continually

have to reinvent curriculum at every crossroads. If the prestige of the College Boards, the Merit Scholarship Program, and similar universal instruments tend to establish an order of excellence that can be translated to all schools as a guide or standard, there is more good than bad in the process. This assumes that local faculties retain the freedom to use the national standards judiciously.

Testing and Vocational Education

So far in this paper, I have sought to review some of the criticisms of testing as I see them. I have tried to respond briefly as a practitioner in the craft of school management to the product that you as test makers have created. At this point I should like to discuss one topic that I have never heard anything about from the critics. It is the whole question of the function of testing in occupational, vocational, and technical education.

One of the enduring curses in our elementary and secondary schools has been the low esteem attached to vocational-technical education. The prestige of the college preparatory program, with its historic aura of excellence by definition, has implied wrongly that the noncollege student is engaged in something that is less than excellent. I ask that those of you who are concerned with the social responsibilities of testing look upon this as a challenge to your ingenuity.

In the minds of young people, the lack of *relevance* now prevailing in much of our high school curriculum is, in part, a source of the discontent and turbulence in big city schools. A child who has little prospect for college sees little reason to apply himself with diligence to an irrelevant and ill-concealed adaptation of a college-entrance spectrum of courses. On the other hand, he could be engaged enthusiastically and relevantly in preparing directly for the world of work. Relatively few young people are so engaged. There was a time, even a few years ago, when there was a place in our society for the unschooled and the unskilled. This time has passed. And society, including parents and some teachers, tends to stigmatize the young person who enters the traditional vocational school, even though he clearly will not be a likely college candidate. We need to make the vocational program important.

There should be a counterpart to the College Entrance Examination Board, with its prestige, motivation, and universality of standards, for

that other 50 percent of young people in middle schools and high schools who probably will not enter college—at least without first having earned a living for a time. A culminating examination should be created with all the strength and quality and prestige that now characterize the College Board examinations. This examination should include, in part, the appropriate academics of a liberalizing curriculum, but it should have as its principal message a measure of the quality of skilled performance in a given occupation that may be expected of the examinee. This would suggest a whole range of crafts, such as plumbing, carpentry, or electricity, and it would necessarily reach into the exciting new opportunities for young people in the swiftly emerging technologies—electronics, the health sciences, the computer, performing arts and fine arts, chemistry, aviation, and many other fields. Given a goal culminating in a rigorous test of his competencies, the reluctant student of today could find a new relevance in school tomorrow.

The program should have very substantial participation from industry and labor to include lively and visible local councils participating in the performance evaluation. Work-study relationships between the young person and certain industries, trades, and commercial enterprises in his immediate environment would be increased and would add to the relevance of his learning.

Upon completion of his in-school curriculum, the student would be awarded a certificate, which because of the universality of the examination's influence and respectability, would give him the freedom to move into appropriate job categories anywhere. Trading upon the illustrious name of CEEB, we might call this new dimension of testing the JEEP (Job Entry Examination Program). The name is not important, but the function is long overdue.

A Decade of Criticism

In summary, we have had now for the past 10 years a lively and healthy condition of responsible criticism. Perhaps Professor Banesh Hoffman is a good example of the critic—earnest, scholarly, and respected. He says that there does not presently exist any generally satisfactory method for evaluating human abilities, that "objective tests are not worthy of a first-rate mind." That was in '67.

As early as 1962 in his book, *Tyranny of Testing*, he strongly

recommended the creation of a national committee to examine this whole business of testing, and as I have mentioned two or three times, there is now such a group at work, and its members have interviewed Mr. Hoffman, along with other distinguished critics.

On the other hand, in this debate about testing are the views of the practitioner, some of which I have mentioned. For example, John Gardner, in his book, *Can We Be Equal and Excellent Too?*, says that:

> Tests are designed to do an unpopular job. It happens that tests are excellent when limited to the uses for which they were designed. Development of standardized tests is one of the great success stories in the objective study of human behavior, although it is now said that tests give an unfair advantage to the privileged individual.

Before tests, says Gardner, many people seriously believed that the less-educated segments of our society were not capable of being educated. He concludes that anyone who attacks the usefulness of tests must suggest workable alternatives.

In preparing my thoughts for this paper, I talked with Lloyd Michaels, who is viewed by most of us in education as the dean of secondary school leaders in America, and who has a distinguished record of leadership at Evanston Township High School in Illinois. When asked about the critics' claims of over-competitive conditions, exaggeration of testing importance, validity, predictability, and so on, Lloyd had this to say: "I do not believe that undue weight is now attached to the testing surrounding college admissions. The test is simply one of several criteria, albeit an important one. School people and admissions officers have learned to use the test with good sense."

This is an important observation from a man whose school possesses all the conditions of a highly competitive population, including racial and socioeconomic differences, and who has been responsible for getting thousands into college over the past 20 or 30 years.

In conclusion, I suppose one might say that testing is something like garbage collection. It finds itself at the mercy of an infinite array of expert observers, as it seeks to serve all the people. It tends to discriminate between the haves and the have-nots. It invades privacy. It rewards conformity—the same size can, keep it in the right place—and it is expected to take on more than was intended—brush, clippings, old mattresses. It is unreliable, unpredictable, and it fails to reach the poor. But if it should stop all of a sudden, there would be the dickens to pay!

Session II

Theme:
The Socially Disadvantaged

Nonschool Variables in the Education of Disadvantaged Children

EDMUND W. GORDON
Teachers College
Columbia University

For several years now, I have been an admirer of the work Dr. Marie Hughes has been doing as an example of one of the kinds of things some of us have been arguing that education ought to be doing for the disadvantaged. In this paper, however, I am going to take a slightly different direction, by no means to demean that effort, but to suggest that there is another dimension that must be considered in the education of disadvantaged youngsters.

The current crisis in the New York City public schools calls dramatically to our attention the importance of factors not traditionally thought to be of much significance in the education of disadvantaged children. From earlier misperceptions of poor and minority-group families as disinterested in education, we have been shoved to the other end of the continuum and now view the problem as one of their wanting to control the education of their children. This politically focused variable is only one of several nonschool variables which deserve serious attention in planning and implementing educational programs for disadvantaged children. Among these are: 1. economic and ethnic integration in the school; 2. political and social determinants of the school's role in the development of children; 3. accountability of the school to the families and communities it serves; 4. pupil characteristics (physical, intellective, emotional, and cultural) and the relevance of educational programs; and 5. perceived opportunity and academically relevant behaviors in disadvantaged children. For institutions concerned with educational testing and assessment, these variables lead to an increased emphasis on the assessment of informal as well as formal learning environments as opposed to a

preoccupation with assessment of learning potential and academic achievement.

It should be of particular interest to educators to note that some of the most important factors influencing efforts at improving educational opportunities for disadvantaged children are outside the control or influence of classroom teachers. This is of particular importance to me since so much of my writing on the subject has been directed at changing what teachers do, at modifying the attitudes they bring to the classroom, and at broadening the role and function of the school. The fact is that recent research and naturally occurring events strongly suggest that the way in which children are grouped in school, and the backgrounds from which they come, are more telling determinants of the outcomes of instruction than are instructional practices and materials. If we look simply at changes in achievement levels for poor or minority-group children, the gains shown in children who have been educated in economically and ethnically integrated schools are more substantial than the gains associated with differences in school factors such as quality of physical plant, quality of teachers, and richness of instructional materials. If these findings are sustained, it appears that the actions of the judiciary (the courts) have been more effective in improving the quality of educational achievement than have our schools.

I am troubled, however, by this conclusion because I am not prepared to believe that quality of education in school factors is less significant than the impact of out-of-school factors. I am more inclined to believe that in-school factors *can be* more powerful influences if schooling can be sufficiently improved. But given the present state of the art, it appears that we know more about what to do outside the classroom than inside to make a difference in academic achievement.

There are differences, however, between what we know how to do and what we are willing to do. The fact that there is so little progress on the school desegregation front and no rush to even mix children by economic group is a case in point. Clearly, the case for integration must be qualified in view of mixed findings on this issue, but the weight of tradition and conviction dictates delay.

Parent Participation

For many years we have accepted as a basic tenet in education the pivotal roles of parents and community in the development of educational services and in the success the school is likely to have with the child. We have used as an excuse for our failure to adequately educate disadvantaged children the alleged lack of interest and involvement in the school and its program of parents and community members. The assumptions about parents' attitudes toward education seem to have been inferred from correlational evidence, since the problem has not been explicitly studied. What we see is that where home and community are actively involved with the school in planning for and educating children, school progress and academic achievement tend to be higher. Where this is not the case, the incidence of school retardation tends to be greater. More active and interested parents are assumed to contribute to more productive pupils or more productive schools. However, we have not adequately dealt with the fact that those schools in which there is more parent and community involvement also are the schools where income and social status are high. To strengthen this assumed relationship between parent participation and school achievement are the findings that such participation is also associated with greater identification with the school, its goals and values. This suggests that the role the school can play in the development of the child is greatly influenced by such factors as social participation and identification. The school becomes a more powerful force if it, as an institution, is viewed by the family and the child as congruent with and contributing to the values and goals of the child.

Accountability of the School

There is not only this assumed social determinant of the school's role but growing recognition of a prominent and maybe even more powerful political determinant. It is argued by some that more important than congruence of values and identification born of participation is the implicit accountability of the school to parents in high income or high status communities. In other words, it is not their greater participation or the greater resources of these schools but the knowledge on the part of the school people that they are accountable for what happens to children in their schools.

By accountability we mean a procedure through which all aspects of the school are subject to evaluation by the community served and through which this evaluation will result in change when deemed necessary. Accountability involves ensuring that there are measures to determine whether or not specified objectives are reached by the children, and that channels for change exist when they are not.

Wilkerson and Gordon have suggested that it be a responsibility of the schools to ensure that all children—with the exception of about 5 percent who are truly mentally defective—reach the level of academic achievement comparable to graduation from a good high school. If the school cannot fulfill this responsibility in the latter half of this century, it shall have ceased to serve society as a socially functional institution. Poverty, poor family background, insufficient environmental supports for learning may mean that radical changes must be made in the processes of instruction; however, these conditions cannot be excuses for the failure to educate.

Adequate systems of accountability have not been developed in either privileged or underprivileged communities. The teachers and the schools have been free to plod along year after year with little systematic attention given to whether satisfactory products are wrought with the human material received. Certainly no school has identified its goal as universal achievement of specified academic goals. In low income and minority communities a mere 15 percent and 20 percent rate of success as early as the third and fourth grades has been acceptable. Protests and professional concern have resulted in some effort to improve the schools' effectiveness. But nowhere have parents been able to say "No!" to the current crop of experimenters or practitioners. Nowhere have they been able to sit down with authority to participate in the review of institutional evaluation data and in the making of decisions about new directions and alternative strategies. Under these conditions the school has been free to conclude that alleged or acknowledged poor results are due to inadequacies in the children and in their homes, and not in the schools and their teachers.

Recently, however, this arrangement has become unacceptable to an increasing number of ghetto communities, where the people have received an endless succession of "innovative" programs, which unsuccessfully engineer their lives and perpetuate low achievement as well as feelings of powerlessness. People in these communities argue that a central condition for meaningful decentralization and significant

improvement in urban public education is that power be established to hold the schools accountable for what they do or fail to do for the children.

Parent participation and accountability of the school to the community, then, have emerged as important sociopolitical factors influencing schooling for the disadvantaged. By withholding participation, parents can possibly defeat or retard the school's function and their children's progress. By preventing participation or denying accountability, it appears that the school weakens its influence on the development of children.

These first three nonschool variables suggest that the school needs to be as concerned with the sociology and politics of educational organization and administration as with the psychology of instruction and learning. Two additional groups of variables specifically referable to the youngsters themselves seem important.

Understanding Students' Backgrounds

A great body of literature has developed on the characteristics of disadvantaged youngsters. Dr. Hughes referred to it, and I shan't attempt to review all of it here; but it is important to recognize, as we point the finger at the school or at the community, that in trying to understand programs for disadvantaged youngsters, it is also important to understand what these youngsters themselves bring to the school. Not only is this an important consideration at the level of those of us who evaluate, but even more important at the level of those who do the actual teaching.

I submit that one of the problems in the education of disadvantaged children has been the insufficient congruence between many of the characteristics and special needs of these children and the manner in which learning experiences and materials are designed. Very little effort has been directed at trying to adapt learning experiences to variations in style, temperament, values, and background experiences. Reduction in the demands made of these children is more usually practiced than are changes in instructional mode and individualization.

If we look at the health status of school children, we find the disadvantaged youngster at the lower end of the continuum with a higher incidence of malnutrition, a higher incidence of neurologic insult, and

a higher incidence of moderate-to-severe debilitating conditions. If we look at the programs that have been designed for such children, we do not see a corresponding higher degree of sophistication in correcting or treating these kinds of conditions. If we look at intellectual function, particularly on our standard measures of intelligence, we note that poor children score low. However, we have not given serious consideration to the additional fact of differential patterning of intellectual function, which may in part account for that lower score—an aspect of our understanding of this problem that is essential to educational planning.

At the level of emotional differences, there is a whole array of negative characteristics associated with such youngsters, most of them relating in one way or another to degree of motivation, degree of task involvement, and interest in academic affairs. Our tendency again has been to regard the characteristics as negative in these youngsters when they may be positive, or at least appropriate, reactions to difficult conditions of life.

If we begin to look at this group of affective characteristics and examine the way in which these youngsters perceive opportunity and consider the relevance of their educational opportunities to their perception of opportunities, the possibly negative affective characteristics appear not only as handicaps to the children but possibly as the kinds of motivators for educational change that the school may need.

For one who perceives the world as essentially hostile to him, as relatively nonsupportive of his efforts, as relatively lacking in a meaningful opportunity for him, there is little motivation to become deeply involved in tasks that take him into that world for further frustration.

In recent years, the job opportunities for these youngsters have risen sharply. It is no longer entirely a question of factors that relate to lack of motivation, lack of involvement, as was earlier assumed, or lack of any perception of job opportunities at the end of the line for them. At the same time that we have begun to increase job opportunities for this segment of the population, however, we have also seen in our more affluent population an increasing disengagement from concern for vocational opportunities, with the opportunity to work as a norm, while there has been increasing preoccupation with what some people call an opportunity to live. I am not sure that this dichotomy between work and life is limited to the affluent segment of our society. I suspect it is equally a part of the thinking of the disadvantaged youngsters.

We have progressed on the work front but we are lagging behind on the front that is possibly the more appropriate goal—opportunities for the designing of that kind of life and opportunities for the pursuit of those kinds of activities of life that are truly sustaining of the spirit rather than primarily sustaining of the body.

This problem may become more difficult in the next few years if we move into what has been anticipated by some people as a possible shift to the political right. It is not just a problem that political conservatives might win. I think most of us are relatively confident that at least the extreme right will not prevail. But, there is also an awareness that the influence of such a strong rightist movement will be strong in pulling the center to the right, thus increasing the chance that opportunities for upward mobility will decrease and that the society and its institutions will regress in some of the areas that now hold promise for the young. These developments will increase the concerns of this youthful group that I have identified as being now considerably less preoccupied with opportunities for work and more preoccupied with opportunities for living.

I suspect it is here that we will find the greatest challenge to the school. Three to five years ago I would have called reorganizing our approach to teaching and learning in school the most urgent task of the public schools. Today, I must say that I see that challenge in a somewhat different light. I think it is helping youngsters find meaning in their school experience and meaning in their lives—the reintroduction of values and humanistic concerns into their education.

The Responsibility of the Community

Educators have good reason to be concerned, sensitive, and even to feel guilty about the extent to which we have failed in our efforts at educating the underprivileged. The fact is that we have done a lousy job, and despite all our breast beating, agitation, and action, we haven't done much to improve our past performance. After reviewing a good bit of what is offered as compensatory education, I have been forced to conclude that we have provided little compensation. We have met with little success. The quantity of effort has been substantial, but the quality leaves much to be hoped for. However, we would be mistaken if we were to conclude that the need is only for the school to change or improve. Certainly it must improve in many areas in

ways that we are able to specify as well as in ways that are presently unrecognized. But there is a larger arena in which the school and the disadvantaged function or malfunction that is a community—a social system that may confront the school and these students with problems beyond their competencies to overcome. To improve their education it may be necessary to modify that social system so that the negatives in their characteristics may be corrected or produced in lesser abundance, and their other characteristics may be more appropriately utilized in their development. It has been noted that "The democratic problem in education is not primarily a problem of training children; it is a problem of making a community in which children cannot help growing up to be democratic, intelligent, disciplined to freedom, reverent to the goods of life—eager to share in the tasks of the age. A school cannot produce this result; nothing but a community can do so."

If, then, we are concerned with improving education for disadvantaged children, and if we are concerned at this conference with the contribution that measurement can make, the clear focus of new work in this field needs to be placed on the assessment of informal as well as formal environments, learning environments, learning transactions, and conditions for learning. To add to our growing sophistication in the assessment of behavior and potential we need to develop better procedures for the assessment of environments and transactions in which, and through which, learning takes place.

It may be that as we understand these factors and manipulate them, we may find that they are more important than the things we are likely to know in the next few years about manipulating the more formal aspects of the classroom.

Issues and Strategies in Employment of the Disadvantaged

ALBERT P. MASLOW
U. S. Civil Service Commission

Employment—a meaningful place in the working society—persists in the minds of many as the number one remedy for most problems of minority groups. The paths toward this goal—and the many obstacles—have been widely discussed. From the difficulties and disappointments of early efforts by government and business there is emerging, I believe, a more honest perception of the nature and stubborn complexity of the challenge.

There are two major factors involved here: One is the shifting structure of jobs in our society; the other is the barrier to penetration of the labor market by minority groups. To the other gaps identified in recent years, we can add the "employability gap." The unskilled, the untrained, the uneducated are falling farther behind. Traditional entry-level jobs are decreasing; educational demands for both old and new occupations are climbing. Thus, simply to hold their own and perhaps to inch ahead, the minority must claim a relatively greater number of semiskilled and skilled jobs, and at a faster rate, than in the past or present (31).

The barriers are rooted not only in inadequate education and discriminatory employment practices; they also rise out of what has been aptly called a "trained incapacity" to enter the world of work (5):

—The Negro, excluded from direct experience or contact with people in high-status jobs, has only "irrelevant" work models.

—He is excluded from the "work ethos." That is, he does not learn the "intrinsic value" of work, and has less faith in the system of rewards.

—He is alienated from "job ways."

Thus, much more is involved than will be settled just by creating jobs and changing our hiring practices.

The wide sweep of these problems contrasts with the quite modest purpose of this paper. That purpose is, simply, to review the main issues from the standpoint of the psychologist and of related professionals who have a role to play in this national effort and to suggest strategies that may help us fulfill that role.

Members of a "good" profession, we are told, guide their practices and policies by a sense of social responsibility (1). As we take part in urgent programs to get people into jobs, we have a responsibility to develop and apply methods that are objective and technically sound, that do not force people into work for which they are ill-suited. So, especially with minority employment, what we do, and why, needs to be communicable to the job seeker as well as to the employer and, no doubt, to the arbitrator (17, 19).

Under pressures to demonstrate fair practices, private employers will, I believe, adopt more of the precepts historic in public merit systems. They will need to make public their decisions not only as to whether a man is qualified, but also as to how well qualified he is in comparison to others, and then to use these relative judgments in a systematic and impartial way in making job offers.

If a single guiding principle can be distilled out of recent experience, it is this: The process of minority employment must be viewed and dealt with as a total system. The Equal Employment Opportunity Commission (4) advocates such a "total assessment system" encompassing job analysis, special recruiting efforts, job-related screening, interviewing, testing, and validation. It also properly stresses the "mutual interdependence" of the components of the system.

One impression from such an analysis—perhaps unintended—is that jobs are relatively fixed, and that the purpose of the total system is to locate those persons in the labor market who best match the job requirements. This point of view rests on several assumptions:

—The job seeker knows or can find out a good deal about the nature and demands of available jobs.

—He can relate his own experience and needs to his perception of the job.

—He has the resources—both material and psychological—to cope with the employment process.

— He can compete and survive under supervisory practices and training programs which are primarily production-oriented and designed for and by employees quite different from himself in many ways.

This approach—finding the right man for the job—may well serve the employer in a classic demand-and-supply sense. However, it is unlikely, as recent experience shows, to meet the needs of disadvantaged individuals for a productive role in our economic life. Furthermore, as Wolfbein (31) points out, if programs for a guaranteed income and employment come to pass, we will be expected to know how, in practical terms, to motivate, to train, and to retain workers in careers, not just in jobs. Otherwise these efforts may well be condemned as make-work.

Jobs for the Disadvantaged

The term "socially disadvantaged," like most labels, diverts attention from individuals and disguises the complex makeup of the groups in need of help. Disadvantaged is not simply being unemployed. It is not simply being a member of a minority ethnic group. It is rather a relative concept to be defined in terms of the individual and his environment. Bloom, Davis, and Hess (3) describe the disadvantaged as those whose problem originates in early experience in the home, where parents failed to "transmit the cultural patterns necessary for the types of learning characteristic of the schools and the larger society," a society which puts a premium on education and the ability to adapt to changes in work and in other aspects of life.

Yet, in any terms, the stark outcome of such deficit is reflected in the labor force picture (13). An estimated 11 million people had some period of unemployment during 1966. Prolonged unemployment—out of work for over half the year—hit over one million of these, more heavily among the young, the old, the unskilled, and the nonwhite. Another large group—difficult to count but perhaps as many as two million—comprises the *under*employed (again, more likely to be nonwhites in part-time service occupations). In September 1967, over five million wanted a job, but were not looking for work. The most disturbing subgroup among these was the three-quarter million no longer seeking jobs because they felt unable to get them.

Scattered through these social morbidity data are the hard-core

unemployed—500,000, mainly teenagers and men over 45 with no job experience or marketable skills or with serious job handicaps in health or education. Most are Negroes, Mexican-Americans, Puerto Ricans, and Indians.

Governmental and private efforts, in many instances under forced draft, have brought help to large numbers of this underclass. Traditional selection practices have been cast aside to get people into productive jobs. Many are found to need mainly additional training and supportive services to get and hold jobs. But as the needs of this group are being met, the hard core emerges even more clearly as the central and unresolved problem.

For the first group, the assumption seems workable that we are dealing with familiar continuous variables in employment but at lower ranges than we have been accustomed to.

With the hard core, however, we may well be facing *qualitative* differences. Therefore, simply to modify traditional training and employment practices is not especially responsive.

A useful insight is found in Fine's (7) analysis of the nature of skills which underlie human performance. *Adaptive* skills are those by which the worker adjusts to the "physical, interpersonal and organizational" conditions of the job. These skills, based on early home and school training, include habits of dress, punctuality, reactions to authority and others, and care of property. *Functional* skills are based on education, job training, and work experience; they are the worker's skills in dealing with "things," "data," and "people." Finally, *specific content* skills derive from specialized technical training and extensive job experience.

Traditionally, employers have set their job requirements and have screened applicants in terms of the functional and specific skills, assuming that the employee will get to work on time, will do what he is told, will work safely, and will fit in pretty well with the existing work force. For the hard core, however, we cannot take even these adaptive skills for granted. For example, in one program, the tardiness of many trainees was traced to the simple fact that 25 percent of them could not tell time!

We need also to know a great deal more about the sociopsychological factors that condition job-seeking, job-finding, and job-holding behavior.

Two exciting explorations in the past few years have shown that manpower management *can* be enriched by understanding these fac-

tors. Sheppard and Belitsky (23) used structured interviews, questionnaires, and projective test methods to measure the job-seeking behavior of jobless blue-collar workers, the relative effectiveness of ways they tried to get jobs, and, most significantly, their achievement values and job-interview anxiety.

Some provocative differences in job-seeking behavior showed up and were found to be related differentially to measures of adjustment, motivation, values, and anxiety. The authors argue that such social-psychological characteristics operate as *determinants* of the job-seeking behavior of these unemployed workers.

A practical technique for measuring the motivation to work would add, especially at entry-level occupations, a big gun to our placement arsenal. Indik (12) has reported a study with this objective. Following Atkinson and McClelland, he defines the motivation to work as the *product* of the motive to work, times the expectancy to work, times the incentive to work. Similarly, the motivation to avoid work is the product of the motive to avoid work times the expectancy to avoid work, times the incentive to avoid work. The "residual behavior potential to work" is then defined as the difference between these two products. In this research, scales to measure each of the six concepts were found to be essentially independent and to correlate with several external indicators such as employment status and training status. As Indik points out, such scales support both systematic research and application of motivation measures in selection and guidance.

In filling federal jobs, we already collect, on a routine basis, some data on the educational, economic. and cultural background of job applicants. (For large-scale operations questions are printed right on the test booklet or other personnel forms, and responses are tabulated by test-scoring or optical-scanning equipment.) The sharper tools now emerging should greatly increase our sophistication in helping the unemployed to seek, find, and keep a job.

What bearing do these matters have on an employment strategy? Motivational and other social-psychological variables work in conjunction with the abilities and skills of job applicants and with availability of suitable jobs. Thus, to repeat a familiar theme, *all* aspects need to be carefully integrated. Moreover, one's self-concept and perception of the job are just as important to the disadvantaged, we must assume, as to anyone else.

But traditional recruiting methods and ways of communicating information about job opportunities and demands do not help

develop these insights. Job titles simply do not inform. If the hard core must depend for their job knowledge on their own experiences or on those of relatives, they will learn only about the low-level, unsatisfying jobs which they and their friends hold. So, if employers really want to *reach out*, they must retool practically every employment procedure: the level and kinds of communication, both written and oral, about job demands and opportunities, the forms to be read and filled out, the whole setting and method of test administration, the environment and conduct of interviews, and so on. This outreach process includes providing transportation, following up with job training, and many other steps that must be taken to reconstruct barriers into pathways. For example:

- We need to do a better job of preparing and displaying occupational information. In particular, the opportunity for the disadvantaged to experience varieties of work activities, through job samples in typical work settings, would give room for informed self-selection to operate.
- There is an acute point to the quip that the want ads are the "colored funnies." Dramatic efforts are needed to make the employment process and likelihood of obtaining a job *credible*.
- If measures of motivation become of practical use, employers will be able to attend more carefully to Vroom's (29) proposition that performance is a function of the product of motivation times ability. It is likely that, for many jobs, the hiring of highly motivated people with marginal ability will be found to yield more than the hiring of high-ability applicants with low motivation to work at that kind of job.
- If many jobs for the disadvantaged call mainly for adaptive skills, then the relative emphasis on training and on motivation increases very greatly.

As we turn to strategies for expanding job and career opportunities, we find that the job structure, the training facilities, and the nature of supervision are inseparable in practice as we try to organize useful work for the disadvantaged.

Two general directions of effort can be discerned. The first is in the reorganization of *present jobs*. The second is in the development of *new careers* which are peculiarly suited and accessible to the disad-

vantaged (6, 16).

It is all too ea y to set up jobs simply by reassembling, from existing jobs, many of the simplest and most meaningless tasks for assignment to the new worker. These "Mickey Mouse" jobs, as the teenagers call them, cannot provide the kind of motivation which is so important in shaping the new employee's attitudes toward work. Just as persons at other occupational levels need to see the ·'. ance and the training value of their work, even more do these new workers for whom training and job success is still to be experienced. Above ·ll, employers must guard against the emergence of occupational ghettos and must map out promotional lines and career patterns. Guidelines for the enlargement of jobs and motivation of employees such as those suggested by Herzberg (11) are certainly relevant at entry levels. Finally, in establishing jobs employers must pay attention to the social structure and the inter-group relationships which, even more than work content, affect the perfo.mance and attitudes of the new worker.

At the level of skilled and technical occupations, job redesign is more familiar. Here, the revision of standard journeyman jobs and the greater use of technicians, for example, have a double advantage—to expand the labor supply i.. .ortage occupations and to provide opportunities for many persons who do not yet have the traditional full technical or professional training. This approach, strongly pressed in the Federal Government (26), leads to such career fields as Economics Assistant, Engineering Technician, Psychology Aide and Technician, and so on. The significant point is that each of these fields is designed as a rewarding career in itself, with opportunity for training, promotion, and growth alongside of, and working with, fully trained professional and scientific personnel. The apprenticeable trades have been somewhat more resistive t. change, but, even there, some inroads are being made.

The redesign of existing work is, however, subject to economic fluctuation and the need for employers to protect their profit margins. In the longer view, perhaps more productive will be setting up *new careers* in providing public services to our nation in health, education, and community betterment. The potential number of such jobs has been estimated at over five million (18). They need not demand a high degree of preparation. To the disadvantaged these jobs would offer the reward of helping to better conditions which they have suffered at first hand.

In all these efforts, the emphasis will of course shift from selection

to training. Thus, we should ask: What precepts in training are responsive to the need? As present jobs at entry levels are restructured, the new employees will need training in adaptive skills. Yet these skills, being nonspecific to job tasks, do not tie workers to a specific job. This, in part, may account for the high turnover experienced at this level. In addition to good work habits, employers need to provide experience in a variety of tasks to broaden the knowledge of the worker about the world of work. Earlier this year, in a conference sponsored by the New York State Psychological Association, it was repeatedly stressed that training must have a positive purpose and a target job that really exists. Academic training and counseling must be provided where necessary. Successful training includes the establishment of immediate rewards as quickly as adaptive and job skills are mastered. The goal of training—developing the ability to meet job-performance standards—must not be relaxed either in fact or in the perception of the trainees. Softness by management leads to mistrust on the part of the trainee and his peers and interferes with his integration into the regular work force.

Obviously, to tamper with job structure is to disturb the enshrined standards of training and experience requirements. Note that by changing requirements one does not simply admit persons with lower degrees of skill. It may well change dramatically the labor supply itself. This will be especially true as barriers based on employability—arrest records, and similar behavioral indices—are pulled down. Furthermore, when redesign of jobs shatters the credentials barrier, the shock waves are quickly felt in the reaction of professional and other protective groups.

Job redesign also means that selection techniques shift from achieved job skills to *trainability*. Similarly, performance criteria and supervisory expectations are altered. Supervisors will most likely put up with considerably less than optimum performance, provided the employee shows willingness to learn and fits in well with the work force. One employer judged an employee satisfactory if he stayed on the job for one year and showed up for work quite regularly!

The demands on *first line* supervisors, always a serious management concern, are especially oppressive. To survive and be effective in this program, the supervisor himself must readjust his expectations of performance, and this, in turn, will affect how he evaluates his employees. Thus, there are a host of subtle problems in supervision that we need to learn more about and carry into our systems for selecting

and training supervisors.

To illustrate: Following racial disorders in San Francisco, an experiment was conducted in which close to 500 men and women were hired as postal clerks without regard to aptitudes, or even to arrest records and other background data (8). Their performance and retention rates over a year were compared with a control group selected in the normal way, and they were tested *after* hiring. Briefly, we found overall retention rates were the same for both groups. Although the control group did somewhat better work, 70 percent of the experimental group who stayed a full year were also fully earning their pay. The aptitude test showed a moderate relation to performance, but a large percent of those with lowest scores were able to learn and perform the work. Of greatest value, however, are the insights we gained for management of such workers. We found that the new worker needs, above all, a stable relationship with his supervisor and his work group. He needs controlled and regular hours of work and clear-cut task assignments. It is clearly disruptive to herd such employees into work gangs with frequent reassignments to different supervisors to handle shifting workloads. Particularly, we learned to guard against frequently assigning them to new supervisors either inadvertently or as a result of marginal performance. Such management practices simply resulted in the employees *most* in need of supervision getting the least.

Our experiences in job design, training, and supervision expose how bare are the traditional practices of job analysis and description. There is a tremendous contribution to be made in finding ways to include social, situational, and motivational variables in job descriptions. Information on the real nature of Negro-white interaction on the job or about the opportunities to satisfy achievement needs, for example, would be immensely more useful than a simple recital of tasks and procedures.

Problems of Employment Testing

Much of the criticism of current employment practices zeroes in on testing. Many believe that current employment tests are biased both because of the way they have been developed in our society, and by misuse. What are the facts? Just four years ago, in a symposium at the American Psychological Association convention, the point was made,

and is still true, that we know only this:

1. Mean scores for Negroes are generally lower than mean scores for whites on employment tests of the general-ability type, despite the considerable overlapping of distributions.
2. There is no research on the *validity* of tests based upon Negro samples to support any generalization whatever (20).

There is a legitimate question as to whether the kinds of tests now in use are, or can be, relevant to the truly disadvantaged. In a review of experimental and demonstration projects under the Manpower Development and Training Act, Gordon (9) found what I can only describe, in charity, as a frivolous use of tests and test data both for selection and counseling. The list comprised primarily "old standard tests" whose selection seemed "more determined by factors of expediency than by a careful and comprehensive evaluation. . . ." From the project data reported, Gordon concludes that the "validity of tests for predicting job success of disadvantaged youth is yet to be established, and the weight of the evidence suggests that such validity will *not* be found."

To find the validity of tests, in practical use, has proved over the years to be singularly difficult. We are all too familiar with the obstacles: small and selected samples, global rather than specific criteria, changes in job demands, and unpredictable postemployment changes in the employee through learning, and other adaptations.

In minority testing the research problem takes on added dimensions. Following Guion (10) we can perhaps agree that "unfair discrimination exists when persons with equal probability of success on the job have *un*equal probability of being hired." This definition is broad enough to apply to the entire process used by an employer. In limited application to testing, it calls for determining the predictive value of tests separately for minority and nonminority groups. Such an effort is beset with serious administrative, statistical, analytical, and political problems (10, 24). For example, relationships for one group might be linear and for another nonlinear. Because of probable difference in variances, the predictive significance of scores at different ranges may differ for the two groups. The criteria can also have different meanings. For example, Ruda and Albright (22) who found race-related differences in turnover commented that "the greater tendency for Negroes than whites to remain on the job simply reflects the easier job mo-

bility for the whites." Furthermore, in the practical situation, whether in fact discrimination does occur will depend upon both where critical scores are set and how scores are used in arranging the job applicants in an order of priority. Thus, even *if* populations of minority and nonminority employees were available, *if* minority status could be openly identified, and *if* dependable criteria were at hand, a clear-cut answer for a particular test would necessarily involve a complex analysis of selection ratios, regression lines, and the linearity of the relationships.

It is small wonder that Tenopyr concluded a review of several widely cited studies by stating that, because of methodological problems, nothing conclusive can yet be said about test bias. "Not only is more research needed but better research is essential."

These design problems are all too real. Over the past several years, Educational Testing Service and the Civil Service Commission have jointly sponsored a study of minority-group testing with support from the Ford Foundation. The first phase of this study was wisely viewed as a feasibility study. We were able to identify a substantial group (160) of Negroes among 460 medical technicians in Veterans Administration hospitals throughout the country. These technicians were given 10 job-related aptitude, achievement, and specific job-knowledge tests. At the same time, background and job-performance data were collected.

Whereas in key background items there were practically no meaningful differences between the Negroes and the non-Negroes, there were, on every test, significant differences in group mean scores. The most intriguing, yet puzzling, data relate to the criterion analysis. There were indications that the reliabilities of ratings were different for the two groups. There emerged a strong relation between supervisor ratings of job knowledge and objective tests of job knowledge when whites rated whites, when whites rated Negroes, and when Negroes rated Negroes. It seemed here that the supervisors based their judgment heavily on the employee's actual job information. However, when Negro supervisors rated white employees, the correlation vanished. What do we make of this? Are Negro supervisors biased against whites? Do they ignore the job knowledge of white employees and base their rating on other factors? Are the less-able whites assigned to the Negro supervisors? I raise these questions simply to make the point that the race of the rater adds one more critical variable to this problem. Some other tentative findings: The best predictive tests are

consistently more valid for Negroes than for non-Negroes, either against ratings of specific performance dimensions or a job-knowledge test. There were some suggestions in the research that differences in test strategy—that is, in the way Negroes approach the testing situation—may have a bearing on their performance.

From this pilot study we have concluded that such research, although very expensive and time consuming, is feasible, but there are many problems of analysis and interpretation yet to be worked through. We hope that this pilot study will be followed by a series of studies with other occupational groups.

The Question of Language in Tests

The question of language in employment tests deserves a special comment. In tests, words are the medium as well as the message. The "poverty of culture" shows up most clearly in language deficit. We find that educational achievement which falls below the fifth or sixth grade amounts to *functional* illiteracy for work. The language demands of a job are not, however, confined to the language needed to carry out specific job tasks. The employee also must be able to read and understand union and company publications and to communicate with others around him. Thus, the true language demands of a job derive from the total job context. A most rewarding field for research I think would be a schema for the analysis and scaling of jobs in terms of language and other communications requirements. This might yield realistic information that is critically needed if we are to help the disadvantaged.

There are several other precautions. Differences in language ability may very well change the meaning of tests for the applicant. What may be a quantitative reasoning item to one applicant may become to another a pure challenge to his reading ability. A particularly troublesome problem is that on many verbal types of tests females will score better than males and embarrass employers whose objective is to hire unemployed young males.

Although I have been speaking of Negroes as the predominant minority group, I do not thereby intend to slight the special concerns of other groups such as Spanish-speaking minorities, Indians, and others. Some argue that it would be more fair to test persons of Spanish-American background in the Spanish language. In our

federal system we encourage their employment by use of Spanish recruiting literature, sample test material in Spanish, and so on. But we have approached somewhat more gingerly the idea of actually presenting test questions both in English and in Spanish. This would provoke broad policy issues with respect to many other language minorities. Moreover, in several studies (15) we found that, for groups educated in United States high schools, competence in written English is very likely to be at least equal to competence in *written* Spanish. For such groups bilingual test material proved of little help. In only a few individual cases where the educational level was low, and the commitment to Spanish in daily living quite high, was some improvement over an initially poor English test score achieved.

As a matter of policy, therefore, we are not at present considering wholesale conversion of tests into English-Spanish. However, findings that apply to the Southwest do not necessarily make sense in Puerto Rico or other areas where education is in a completely Spanish environment.

To sum up: The question—is this test biased?—may be almost impossible to answer. Yet a few guidelines may reduce some sources of obvious discrimination:

1. When tests in fact have *no* validity for either group, whites will generally have an unfair advantage over Negroes in any personnel action which selects from the top. This is simply because whites tend to score higher. (Specific education requirements, or length-of-experience requirements, may work in the same way.) Our task is to identify these elements and eliminate or modify the impact of the predictors.

2. When a test has about equally high validities for both groups, when Negroes score lower, and when the spread of scores is the same or less for Negroes than for whites, the result is that (in merit-system situations) less-able whites are selected *ahead* of able Negroes. As selection proceeds down the score range, the Negro may be placed at a further relative disadvantage because he may be increasingly outnumbered by the whites at any point on the distribution above the cutoff score.

3. A special case arises when recruiting is continuous and the employer is, in effect, continually combing the labor market and placing the best candidates at the top of the list. In this case minority groups would tend to cluster at the lower reaches of employment

lists while the jobs are filled by the continuing inflow of higher scoring majority candidates. This kind of situation gives rise to demands that employers establish lists of qualified people, then close off competition and utilize the entire original "acceptable" list before going again into the open labor market.

4. A crucial decision is whether and where to set a cutoff score:

   a) Again, with typical mean score and variance differences, a high cutoff score will effectively lock out the minorities. It is justified only by strong evidence of equal validities and score distributions.

   b) In the absence of validity data, cutoff scores should be set no higher than the point below which experience or content analysis or other data indicate failure is quite likely. And if the original deficit could be readily overcome by training, the use of any cutoff score would be hard to defend.

   c) In general, therefore, in the absence of data showing a linear relationship, scores should not be used to array applicants; order of selection should be based on other job-related factors.

5. Precautions need be taken in the construction and administration of tests. Adequate pretest material, a supportive, tension-reducing atmosphere, and steps to ensure that the applicant knows exactly what to do provide the necessary structure for fair testing.

As the failure of classical predictive validity becomes too obvious to deny, attention is shifting to synthetic validation (2). This involves three steps: 1) analysis of the job elements—abilities, skills, personal characteristics—that are required for effective performance; 2) selection of tests and other instruments known to be acceptable measures of these elements; and 3) combination of these measures into a single decision strategy.

About 15 years ago, E. S. Primoff (21) of our staff developed the J-Coefficient for the selection of aptitude tests (30, 14). The J-Coefficient expresses the degree of agreement between elements in a job and in a test. It is statistically equivalent to a correlation coefficient. Workers and supervisors who know the job use a frequency-scaling method and special rating forms to identify important job elements. Test values for elements are estimated by test specialists and stabilized over time by validation studies. This test-selection method compares very favorably with other methods of assembling test batteries (25).

Job titles may be infinite, but job elements are not. We have found that for trades and industrial jobs about 90 elements cover most significant job requirements which can be reliably rated. These elements are not independent in a factorial sense, but they do have psychological reality. Especially, the test values have led to many new insights about the real elements in tests and job performance. In literally hundreds of applications, the J-Coefficient method of test selection has proved very economical and effective. Thus, we now select and use tests with confidence in many situations where empirical validity studies cannot initially be conducted.

A Job-element Strategy for Staffing

Beyond test selection, this concept has broadened into a general system that accommodates every variety of personnel appraisal that will contribute to the measurement of critical job elements. In practice, we define a job element and identify four levels of ability in the element. Next we determine the kinds of experience, training and/or test performance which are acceptable evidence for each point on the scale. Then, in evaluating an applicant on this element, we make use of all relevant data—test scores, reference checks, experience and training records, and interview protocols—to reach a judgment as to where he falls on the scale. After he is evaluated on each element in turn, we make a final "whole man" appraisal as to his potential or achieved ability for that job as an individual. Tests are only one of several alternate sources of evidence about an element. For the inexperienced, an aptitude test score might be relevant; for the experienced, a record of his actual job proficiency would be most meaningful, regardless of test score. So we have great flexibility in the range of information used to reach judgments about the applicant. What we are really after is a reliable judgment about his abilities, regardless of how he developed them.

We have reached the point where this general approach is now applied to the staffing of about 800,000 trades and industrial jobs across the nation. Here in the job-element system is a strategy, then, that responds to the needs I outlined earlier in employment of the disadvantaged:

—The focus on observable abilities, rather than job titles, permits

communication with applicants in language they can understand.

—The clear-cut definition of elements simplifies and eases, yet unifies, every step of the employment process.

—Tests and experience or training requirements no longer lock out applicants arbitrarily. Motivational aspects can be given full and systematic weight.

Beyond employment, this approach also offers a psychological rationale for:

—identifying training needs and setting goals;

—grouping jobs into meaningful families and thus helping in job mobility and reassignment;

—arraying jobs into career ladders based on progression of abilities;

—redesigning jobs to fit the abilities of workers;

—appraising performance on the basis of critical abilities.

The precepts of this job-element system have turned out to be most valuable in expanding minority job opportunities. For example, earlier this year, the Civil Service Commission (28) worked in concert with federal agencies to set up worker-trainee jobs, with special recruiting, placement, and training as parts of the program. Typical jobs covered by this plan are fireman-laborer, laundry worker, clerical helper, mess attendant, and custodial laborer. In working out the assessment schema, we confronted the fact that applicants who would rank high on traditional skill-oriented selection measures would be least likely to be satisfied or to stay on. On the other hand, many who would welcome these jobs, as an entry to the working world, would not survive traditional hurdles. So in our examining plan we try to estimate the elements of motivation, interest, and readiness of applicants to work *at this level.* Thus, an applicant would be best qualified —that is, in the top category on a four-point scale—if he showed, for example, evidence of efforts to get work, to take advantage of training under a Manpower Development and Training project or similar training, or had such limited education as to be practically unemployable at higher-level jobs in his community.

At the other end of the scale, those least qualified for these jobs are applicants who are college-bound, who have worked at higher-level

jobs and at higher pay rates, or who are in training for higher-level occupations.

To give effect to this plan, the publicity states in plain and honest English the work to be done and the physical demands. Simplified forms are used to get information on schooling and work experience, on willingness to do special tasks, and on job interests and expectations. Where necessary, these forms are filled out by the recruiter in an interview with the applicant.

In a few months, over 600 people have been hired. Early reports tell us that these new workers are doing well.

The concept of a ceiling in job standards has been a hard sell, especially in public personnel settings. One reason is that there is a widespread implicit assumption of linearity in the relation of predictors to job performance. Also many feel that every citizen is entitled to a job which may be well below his skill or needs if he simply insists on being considered.

I believe the job-element rationale offers a reasonable and acceptable way to deal with this issue, as well as with others that affect minority employment.

In my opening comments, I cited our responsibility to apply professionally sound methods to these issues. This standard is necessary, but not sufficient; the vital ingredient to real progress toward equality of employment opportunity is the *commitment*, on the part of all of us, to make it a reality.

REFERENCES

1. American Psychological Association. Psychology as a profession. Washington, D. C.: 1968.
2. Balma, J. *et al.* The development of processes for indirect or synthetic validity (a symposium). *Personnel Psychology*, 1959, 12, No. 3, 395-420.
3. Bloom, B. S., Davis, A., & Hess, R. *Compensatory education for cultural deprivation.* New York: Holt, Rinehart & Winston, 1965.
4. Equal Employment Opportunity Commission. Guidelines on employment testing procedures. Washington, D. C.: 1966.
5. Ferman, L., Kornbluh, J., & Miller, J. A. *Negroes and jobs.* Ann Arbor, Michigan: University of Michigan Press, 1968.

6. Fine, S. A. *Guidelines for the design of new careers.* Kalamazoo, Michigan: W. E. Upjohn Institute for Employment Research, 1967.

7. Fine, S. A. Nature of skill: implications for education and training. Washington, D. C.: Proceedings, 75th Annual Convention, American Psychological Association. 365-6, 1967.

8. Futransky, D. L. & Wagner, D. On the job follow-up of postal clerks hired in San Francisco without employment tests. Washington, D. C.: U. S. Civil Service Commission, 1968.

9. Gordon, J. E. Testing, counselling, and supportive services for disadvantaged youth. Washington, D. C.: U. S. Department of Labor, 1968.

10. Guion, R. M. Employment tests and discriminatory hiring. *Industrial Relations*, 1966, 5, No. 2, 20-37.

11. Herzberg, F. One more time: how do you motivate employees? *Harvard Business Review*, 1968, 46, No. 1, 53-62.

12. Indik, B. P. *The motivation to work.* New Brunswick, N. J.: Rutgers University, Institute of Management and Labor Relations, 1966.

13. Manpower Report of the President. Washington, D. C.: U. S. Government Printing Office, 1968.

14. Maslow, A. P. Job analysis techniques. Paper presented at Public Personnel Association Institute, Chicago: 1958.

15. Maslow, A. P. & Futransky, D. L. Effect on test score of presenting verbal test questions in an English-Spanish format to a predominantly Spanish-American group. Washington, D. C.: U. S. Civil Service Commission, 1968.

16. McPherron, D. R. Jobs for the disadvantaged—how? *Public Personnel Review*, 1967, 28, No. 3, 165-168.

17. Metzler, J. H. & Kohrs, E. V. Tests and the requirements of the job. *The Arbitration Journal*, 1965, No. 2, 103-111.

18. National Commission on Technology, Automation, and Economic Progress. *Technology and the American economy.* Washington, D. C.: U. S. Government Printing Office, 1966.

19. Newell, R. Psychological testing and collective bargaining. *The American Federationist*, 1968. 75, No. 2, 18-23.

20. Parrish, J. A. *et al.* The industrial psychologist: selection and equal employment opportunity (a symposium). *Personnel Psychology*, 1966. 19, 1-39.

21. Primoff, E. S. The J-Coefficient approach to jobs and tests. *Personnel Administration*, 1957. 20, No. 3, 34-40.

22. Ruda, E. & Albright, L. E. Racial differences in selection instruments related to subsequent job performance. *Personnel Psychology*, 1968. 21, No. 1, 31-41.

23. Sheppard, H. L. & Belitsky, A. H. *The job hunt*. Baltimore: The Johns Hopkins Press, 1966.

24. Tenopyr, M. L. Personnel testing and fair employment practices. Paper presented at meeting of Western Psychological Association, San Diego: March 1968.

25. Trattner, M. H. Comparison of three methods for assembling aptitude test batteries. *Personnel Psychology*, 1963. 16, No. 3, 221-232.

26. U. S. Civil Service Commission. *Carrying out operation* MUST. Bulletin No. 300-11. Washington, D. C.: December 23, 1966.

27. U. S. Civil Service Commission. *Job element examining* (Handbook X-118C). Washington, D. C.: 1967.

28. U. S. Civil Service Commission. *Examining for jobs at the lowest levels.* Bulletin No. 300-18. Washington, D. C.: April 29, 1968.

29. Vroom, V. H. *Work and motivation*, New York: John Wiley & Sons, 1964.

30. Wherry, R. J. *A review of the j-coefficient*. Washington, D. C.: U. S. Civil Service Commission, 1955.

31. Wolfbein, S. L. *Education and training for full employment*, New York: Columbia University Press, 1967.