Listening. Learning. Leading.®

ETS R&D Scientific and Policy Contributions Series
ETS SPC–11-01

Evaluating Educational Programs

Samuel Ball

April 2011

Evaluating Educational Programs

Samuel Ball

ETS Research Report No. RR-11-15

April 2011

This paper originally was published by Educational Testing Service in 1979.

Series Foreword

Since its founding in 1947, ETS has conducted a significant and wide-ranging research program that has focused on, among other things, psychometric and statistical methodology; educational evaluation; performance assessment and scoring; large-scale assessment and evaluation; cognitive, developmental, personality, and social psychology; and education policy. This broad-based research program has helped build the science and practice of educational measurement, as well as inform policy debates.

In 2010, we began to synthesize these scientific and policy contributions, with the intention to release a series of reports sequentially over the course of 2011 and 2012.

This paper, written by Samuel Ball, was originally published in 1979. It is the only reissue in the series but is a fitting inaugural: The report documents the vigorous program of evaluation research conducted at ETS in the 1960s and 1970s, which helped lay the foundation for this fledgling field.

The report's author, Samuel Ball, was an ETS scientist from 1968-1978, and subsequently Pro-Vice Chancellor at the University of Sydney and Chief Executive Officer with the Board of Studies in the state of Victoria, Australia. Sadly, he passed away unexpectedly in December 2009.

Future reports in the series will focus on each of the major areas of research and education policy in which ETS has played a role.

Ida Lawrence
Senior Vice-President
Research & Development Division
Educational Testing Service

Evaluating Educational Programs

Foreword

The formal evaluation of educational programs is a relatively recent phenomenon, and Educational Testing Service research scientists have been among those striving to chart the unknown waters during the past 15 years. This report is an attempt to record our experiences and the insights we have gained. Clearly, ETS has not done it all; we have learned much from the efforts of others in the educational community as well as from our own endeavors.

It is fitting that the report bears the name of Samuel Ball, for he was one of our most active program evaluators for 10 years and directed several pacesetting studies. He resigned his position as a Senior Research Psychologist in 1978 to accept a chair in education at the University of Sydney in his native Australia.

Dr. Ball still collaborates with us on occasional projects, and it was during a visit earlier this year that he put the finishing touches to this report. We publish it now in the hope that what we have learned about program evaluation will be of value to others in education.

Samuel J. Messick
Vice President for Research
June 1, 1979

An Emerging Profession

Evaluating educational programs is an emerging profession, and Educational Testing Service has played an active role in its development over the past 15 years. The term "program evaluation" only came into wide use in the mid-60s, when efforts at systematically assessing programs multiplied. The purpose of this kind of evaluation is to provide information to decision-makers who have responsibility for existing or proposed educational programs. For instance, program evaluation may be used to help make decisions concerning whether to develop a program (*needs assessment*), how best to develop a program (*formative evaluation*), and whether to modify—or even continue—an existing program (*summative evaluation*).

Needs assessment is the process by which one identifies needs and decides upon priorities among them. *Formative evaluation* refers to the process involved when the evaluator helps the program developer—by pre-testing program materials, for example. *Summative evaluation* is the evaluation of the program after it is in operation. Arguments are rife among program evaluators about what kinds of information should be provided in each of these forms of evaluation.

In general the ETS posture has been to try to obtain the best—that is, the most relevant, valid, and reliable—information that can be obtained within the constraints of cost and time and the needs of the various audiences for the evaluation. Sometimes, this means a tight experimental design with a national sample; at other times, the best information might be obtained through an intensive case study of a single institution. ETS has carried out both traditional and innovative evaluations of both traditional and innovative programs, and staff members also have cooperated with other institutions in planning or executing some aspects of evaluation studies. Along the way, the work by ETS has helped to develop new viewpoints, techniques, and skills.

The Range of ETS Program Evaluation Activities

Program evaluation calls for a wide range of skills, and evaluators come from a variety of disciplines: educational psychology, developmental psychology, psychometrics, sociology, statistics, anthropology, educational administration, and a host of subject-matter areas. As program evaluation began to emerge as a professional concern in these fields, so ETS changed, both structurally and functionally. The structural changes were not exclusively tuned to the needs of conducting program evaluations. Rather, program evaluation, like the teaching of English in a well-run high school, became to some degree the concern of virtually all the professional staff. Thus, new research groups were added, and they augmented the organization's capability to conduct program evaluations.

The functional response was many-faceted. Two of the earliest evaluation studies conducted by ETS indicate the breadth of the range of interest. In 1965, collaborating with the Pennsylvania State Department of Education, Henry Dyer of ETS set out to establish a set of educational goals against which later the performance of the state's educational system could be evaluated. A unique aspect of this endeavor was Dyer's insistence that the goal-setting process be opened up to strong participation by the state's citizens and not left solely to a professional or political elite. (In fact, ETS program evaluation has been marked by a strong emphasis, when at all appropriate, on obtaining community participation.)

The other early evaluation study in which ETS was involved was the now famous Coleman Report (*Equality of Educational Opportunity*), issued in 1966. ETS staff, under the direction of Albert E. Beaton, had major responsibility for design of the study and analysis of the massive data generated. Until then, studies of the effectiveness of the nation's school, especially with respect to programs' educational impact on minorities, had been small-scale. So the collection and analysis of data concerning tens of thousands of students and hundreds of schools and their communities were new experiences for ETS and for the profession of program evaluation.

In the intervening years, the Coleman Report and the Pennsylvania Goals Study have become classics of their kind, and from these two auspicious early efforts, ETS has become a center of major program evaluation. Areas of extensive endeavors have been, and are, diverse. They include computer-aided instruction, aesthetics and creativity in education, educational television, educational programs for prison inmates, reading programs, camping programs, career education, bilingual education, higher education, preschool programs, special education, and drug programs. (For brief descriptions of ETS work in these areas, see Appendix A.) ETS also has evaluated programs relating to year-round schooling, English as a second language, desegregation, performance contracting, women's education, busing, Title I of the Elementary and Secondary Education Act (ESEA), accountability, and basic information systems.

One piece of work which must be mentioned is the *Encyclopedia of Educational Evaluation*, published in 1975 by Jossey-Bass Publishers, Inc. It was edited by Scarvia B. Anderson, Samuel Ball, and Richard T. Murphy, and contains articles by them and 36 other members of the ETS staff. Subtitled *Concepts and Techniques for Evaluating Education and Training Programs*, it contains 141 articles in all.

ETS Contributions to Program Evaluation

Given the innovativeness of many of the programs evaluated, the newness of the profession of program evaluation, and the level of expertise of the ETS staff who have directed these studies, it is not surprising that the evaluations themselves have been marked by innovations for the profession of program evaluation. At the same time, ETS has adopted several principles relative to each aspect of program evaluation. It will be useful to examine these innovations and principles in terms of the phases that a program evaluation usually attends to—goal setting, measurement selection, implementation in the field setting, analysis, and interpretation and presentation of evidence.

Making Goals Explicit

It would be a pleasure to report that virtually every educational program has a well-thought-through set of goals, but it is not so. It is, therefore, necessary at times for program evaluators to help verbalize and clarify the goals of a program to ensure that they are, at least, explicit. Further, the evaluator may even be given goal development as a primary task, as in the Pennsylvania Goals Study. This was seen again in a similar program, when Robert Feldmesser, in 1973, helped the New Jersey State Board of Education establish goals that underwrite conceptually that state's "thorough and efficient" education program.

Work by ETS staff indicates there are four important principles with respect to program goal development and explication. The first of these principles:

> What program developers say their program goals are may bear only a passing resemblance to what the program in fact seems to be doing.

This principle—the occasional surrealistic quality of program goals—has been noted on a number of occasions: for example, assessment instruments developed for a program evaluation on the basis of the stated goals sometimes do not seem at all sensitive to the actual curriculum. As a result, ETS program evaluators seek, whenever possible, to cooperate with program developers to help fashion the goals statement. The evaluators also will attempt to describe the program in operation and relate that description to the stated goals, as in the case of the 1971 evaluation of the second year of *Sesame Street* for Children's Television Workshop by Gerry Ann Bogatz and Samuel Ball. This comparison is an important part of the process and represents sometimes crucial information for decision-makers concerned with developing or modifying a program.

The second principle:

> When program evaluators work cooperatively with developers in making program goals explicit, both the program and the evaluation seem to benefit.

The original *Sesame Street* evaluation in 1970 exemplified the usefulness of this cooperation. At the earliest planning sessions for the program, before it had a name and before it was fully funded, the developers, aided by ETS, hammered out the program goals. Thus, ETS was able to learn at the outset what the program developers had in mind, ensuring sufficient time to provide adequately developed measurement instruments. If the evaluation team had had to wait until the program itself was developed, there would not have been sufficient time to develop the instruments; more important, the evaluators might not have had sufficient understanding of the intended goals—thereby making sensible evaluation unlikely.

The third principle:

> There is often a great deal of empirical research to be conducted before program goals can be specified.

Sometimes, even before goals can be established or a program developed, it is necessary, through empirical research, to indicate that there is a need for the program. An illustration is provided by the 1976 research of Ruth Ekstrom and Marlaine Lockheed into the competencies gained by women through volunteer work and homemaking. The ETS researchers argued that it is desirable for women to resume their education if they wish to after years of absence. But what competencies have they picked up in the

interim that might be worthy of academic credit? By identifying, surveying, and interviewing women who wished to return to formal education, Ekstrom and Lockheed established that many had indeed learned valuable skills and knowledge. Colleges were alerted and some have begun to give credit where credit is due.

Similarly, when the federal government decided to make a concerted attack on the reading problem as it affects the total population, one area of concern was adult reading. But there was little knowledge about it. Was there an adult literacy problem? Could adults read with sufficient understanding such items as newspaper employment advertisements, shopping and movie advertisements, and bus schedules? And in investigating adult literacy, what characterized the reading tasks that should be taken into account? Murphy, in a 1973 study, considered these factors: the *importance* of a task (the need to be able to read the material if only once a year as with income tax forms and instructions), the *intensity* of the task (a person who wants to work in the shipping department will have to read the shipping schedule each day), or the *extensivity* of the task (70 percent of the adult population read a newspaper but it can usually be ignored without gross problems arising). Murphy and other ETS researchers conducted surveys of reading habits and abilities, and this assessment of needs provided the government with information needed to decide on goals and develop appropriate programs.

Still a different kind of needs assessment was conducted by ETS researchers with respect to a school for learning disabled students in 1976. The school catered to children aged 5-18 and had four separate programs and sites. ETS first served as a catalyst, helping the school's staff develop a listing of problems. Then ETS acted as an *amicus curiae*, drawing attention to those problems, making explicit and public what might have been unsaid for want of an appropriate forum. Solving these problems was the purpose of stating new institutional goals—goals that might never have been formally recognized if ETS had not worked with the school to make its needs explicit.

The fourth principle:

> The program evaluator should be conscious of what interested in the unintended outcomes of programs as well as the intended outcomes specified in the program's goal statement.

In program evaluation, the importance of looking for side effects, especially negative ones, has to be considered against the need to put a major effort into assessing progress toward intended outcomes. Often, in this phase of evaluation, the varying interests of evaluators, developers, and funders intersect—and professional, financial, and political considerations are all at odds. At such times, program evaluation becomes as much an art form as an exercise in social science.

A number of articles has been written about this problem by Samuel J. Messick, ETS vice president for research. His viewpoint—the importance of the medical model—has been illustrated in various ETS evaluation studies. His major thesis is that the medical model of program evaluation explicitly recognizes that "…prescriptions for treatment and the evaluation of their effectiveness should take into account not only reported symptoms but other characteristics of the organism and its ecology as well" (Messick, 1975, p. 245). As Messick goes on to point out, this is a call for a systems analysis approach to program evaluation—dealing empirically with the interrelatedness of all the factors and monitoring all outcomes, not just the intended ones.

When, for example, ETS evaluated the first two years of *Sesame Street*, there was obviously pressure to ascertain whether the intended goals of that show were being attained. It was nonetheless possible to look for some of the more likely unintended outcomes: whether the show had negative effects

on heavy viewers going off to kindergarten, and whether the show was achieving impacts in attitudinal areas.

In summative evaluations, to study unintended outcomes is bound to cost more money than to ignore them. It is often difficult to secure increased funding for this purpose. For educational programs with potential national applications, however, ETS strongly supports this more comprehensive approach.

Measuring Program Impact

The letters "ETS" have become almost synonymous in some circles with standardized testing of student achievement. In its program evaluations, ETS naturally uses such tests as appropriate, but frequently the standardized tests are not appropriate measures. In some evaluations, ETS uses both standardized and domain-referenced tests. An example may be seen in *The Electric Company* evaluations of 1973 and 1974 (Ball, Bogatz, K.M. Kazarow, Donald B. Rubin). This televised series, which was intended to teach reading skills to first through fourth graders, was evaluated in some 600 classrooms. One question that was asked during the process concerned the interaction of the student's level of reading attainment and the effectiveness of viewing the series. Do "good" readers learn more from the series than poor readers? So standardized, norm-referenced reading tests were administered, and the students in each grade were divided into deciles on this basis, thereby yielding 10 levels of reading attainment.

Data on the outcomes using the domain-referenced tests were subsequently analyzed for each decile ranking. Thus, ETS was able to specify for what level of reading attainment, in each grade, the series was working best. This kind of conclusion would not have been possible if a specially designed domain-referenced reading test with no external referent had been the only one used, nor if a standardized test, not sensitive to the program's impact, had been the only one used.

Without denying the usefulness of previously designed and developed measures, ETS evaluators have frequently preferred to develop or adapt instruments that would be specifically sensitive to the tasks at hand. Sometimes this measurement effort is carried out in anticipation of the needs of program evaluators for a particular instrument, and sometimes because a current program evaluation requires immediate instrumentation.

An example of the former is the 1976 study of doctoral programs by Mary Jo Clark, Rodney T. Hartnett, and Leonard L. Baird. Existing instruments had been based on surveys in which practitioners in a given discipline were asked to rate the quality of doctoral programs in that discipline. Instead of this "reputational survey" approach, the ETS team developed an array of criteria (e.g., faculty quality, student body quality, resources, academic offerings, alumni performance), all open to objective assessment. This new assessment tool can now be used to assess changes in the quality of the doctoral programs offered by major universities.

Similarly, the 1976 development by ETS of the Kit of Reference Tests for Cognitive Factors (Ekstrom, John French, Harry H. Harman) also provided a tool—one that could be used when evaluating the cognitive structures of teachers or students if these structures were of interest in a particular evaluation. A clearly useful application was in the California study of teaching performance, also in 1976, by Frederick McDonald and Patricia Elias. Teachers with certain kinds of cognitive structures were seen to have differential impacts on student achievement. In the Trismen study of the aesthetics program previously referred to, the factor kit was used to see whether cognitive structures interacted with aesthetic judgments.

Developing special instruments. Examples of the development of specific instrumentation for ETS program evaluations are numerous. Virtually every program evaluation involves, at the very least, some adapting of existing instruments. For example, a questionnaire or interview may be adapted from ones developed for earlier studies. Typically, however, new instruments, including goal-specific tests, are prepared. Some ingenious examples, based on the 1966 work of E.J. Webb, D. F. Campbell, R.D. Schwartz, and L. Sechrest, were suggested by Anderson for evaluating museum programs, and the title of her 1968 article gives a flavor of the unobtrusive measures illustrated—"Noseprints on the Glass."

Another example of ingenuity is Donald A. Trismen's use of 35mm slides as stimuli in the assessment battery of the Education through Vision program. Each slide presented an art masterpiece, and the response options were four abstract designs varying in color. The instruction to the student was to pick the design that best illustrated the masterpiece's coloring.

Using multiple measures. When ETS evaluators have to assess a variable and the usual measures have rather high levels of error inherent in them, they usually resort to "triangulation." That is, they use multiple measures of the same construct, knowing that each measure suffers from a specific weakness. Thus, in 1975, Donald E. Powers evaluated for the Philadelphia school system the impact of dual-audio television—a television show telecast at the same time as a designated FM radio station provided an appropriate educational commentary. One problem in measurement was assessing the amount of contact the student had with the dual-audio television treatment. Powers used home telephone interviews, student questionnaires, and very simple knowledge tests of the characters in the shows to assess whether students had in fact been exposed to the treatment. Each of these three measures has problems associated with it, but the combination provided a useful assessment index.

In some circumstances, ETS evaluators are able to develop measurement techniques that are an integral part of the treatment itself. This unobtrusiveness has clear benefits and is most readily attainable with computer-aided instructional (CAI) programs. Thus, for example, Donald L. Alderman, in the evaluation of TICCIT (a CAI program developed by the Mitre Corporation), obtained for each student such indices as the number of lessons passed, the time spent on line, the number of errors made, and the kinds of errors. And he did this simply by programming the computer to save this information over given periods of time.

Working in Field Settings

Measurement problems cannot be addressed satisfactorily if the setting in which the measures are to be administered is ignored. One of the clear lessons learned in ETS program evaluation studies is that measurement in field settings (home, school, community) poses different problems from measurement conducted in a laboratory.

Program evaluation, ether formative or summative, demands that its empirical elements usually be conducted in natural field settings rather than in more contrived settings, such as a laboratory. Nonetheless, the problems of working in field settings are rarely systematically discussed or researched. In 1975, in an article in the *Encyclopedia of Educational Evaluation*, Bogatz detailed these major aspects:

- Obtaining permission to collect data at a site
- Selecting a field staff
- Training the staff
- Maintaining family/community support

Of course, all the aspects discussed by Bogatz interact with the measurement and design of the program evaluation. A great source of information concerning field operations is the ETS Head Start Longitudinal Study of Disadvantaged Children, directed by Virginia Shipman. Although not primarily a program evaluation, it certainly has generated implications for early childhood programs. It was longitudinal, comprehensive in scope, and large in size, encompassing four sites and, initially, some 2,000 preschoolers. It was clear from the outset that close community ties were essential if only for expediency—although, of course, more important ethical principles were involved. This close relationship with the communities in which the study was conducted involved using local residents as supervisors and testers, establishing local advisory committees, and thus ensuring free, two-way communication between the research team and the community.

The *Sesame Street* evaluation also adopted this approach. In part because of time pressures and in part to ensure valid test results, the ETS evaluators especially developed the tests so that community members with minimal educational attainments could be trained quickly to administer them with proper skill.

Establishing community rapport. In evaluations of street academies by Ronald L. Flaugher, and of education programs in prisons by Flaugher and Samuel Barnett, it was argued that one of the most important elements in successful field relationships is the time an evaluator spends getting to know the interests and concerns of various groups, and lowering barriers of suspicion that frequently separate the educated evaluator and the less-educated program participants. This may not seem a particularly sophisticated or complex point, but many program evaluations have floundered because of an evaluator's lack of regard for disadvantaged communities. Therefore, a firm principle underlying ETS program evaluation is to be concerned with the communities that provide the contexts for the programs being evaluated. Establishing two-way lines of communication with these communities and using community resources whenever possible help ensure a valid evaluation.

Even with the best possible community support, field settings cause problems for measurement. Raymond G. Wasdyke and Jerilee Grandy showed this to be true in a 1976 evaluation when the field setting was literally that—a field setting. In studying the impact of a camping program on New York City grade school pupils, they recognized the need, common to most evaluations, to describe the treatment—in this case the camping experience. Therefore, ETS sent an observer to the campsite with the treatment groups. This person, who was herself skilled in camping, managed not to be an obtrusive participant by maintaining a relatively low profile.

Of course, the problems of the observer can be just as difficult in formal institutions as on the campground. In their 1974 evaluation of Open University materials, Harnett, Clark, Feldmesser, et al. found, as have program evaluators in almost every situation, that there was some defensiveness in each of the institutions where they worked. Both personal and professional contacts were used to allay suspicions. There also was emphasis on an evaluation design that took into account each institution's values. That is, part of the evaluation was specific to the institution, but some common elements across institutions were retained. Thus strategy underscored the evaluators' realization that each institution was different, but allowed ETS to study certain variables across all three participating institutions.

Breaking down the barriers in a field setting is one of the important elements of a successful evaluation, yet each situation demands somewhat different evaluator responses.

Involving program staff. Another way of ensuring that evaluation field staff are accepted by program staff is to make the program staff active participants in the evaluation process. While this is obviously a

technique to be strongly recommended in formative evaluations, it can also be used in summative evaluations. In his 1977 evaluation of PLATO in junior colleges, Murphy could not afford to become the victim of a program developer's fear of an insensitive evaluator. He overcame this potential problem by enlisting the active participation of the junior college and program development staffs. One of Murphy's concerns was that there is no common course across colleges. Introduction to Psychology, for example, might be taught virtually everywhere, but the content can change remarkably, depending on such factors as who teaches the course, where it is taught, and what text is used. Murphy understood this variability and his evaluation of PLATO reflected his concern. It also necessitated considerable input and cooperation from program developers and college teachers working in concert—with Murphy acting as the conductor.

Analyzing the Data

After the principles and strategies used by program evaluators in their field operations are successful and data are obtained, there remains the important phase of data analysis. In practice, of course, the program evaluator thinks through the question of data analysis *before* entering the data collection phase. Plans for analysis help determine what measures to develop, what data to collect, and even, to some extent, how the field operation is to be conducted. Nonetheless, analysis plans drawn up early in the program evaluation cannot remain quite as immutable as the Mosaic Law. To illustrate the need for flexibility, it is useful to turn once again to the heuristic ETS evaluation of *Sesame Street*.

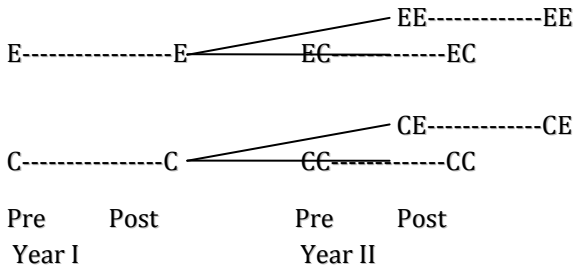
As initially planned, the design of the *Sesame Street* evaluation was a true experiment. The analyses called for were multivariate analyses of covariance, using pretest scores as the covariate. At each site, a pool of eligible preschoolers was obtained by community census, and experimental and control groups were formed by random assignment from these pools. The evaluators were somewhat concerned that those designated to be the experimental (viewing) group might not view—it was a new show on public television, a loose network of TV stations not noted for high viewership. Some members of the *Sesame Street* national research advisory committee counseled ETS to consider paying the experimental group to view. The suggestion was resisted, however, because any efforts above mild and occasional verbal encouragement to view the show would compromise the results. If the experimental group members were paid, and if they then viewed extensively and outperformed the control group at posttest, would the improved performance be due to the viewing, the payment, or some interaction of payment and viewing? Of course, this nice argument proved to be not much more than an exercise in modern scholasticism. In fact, the problem lay not in the treatment group but in the uninformed and unencouraged-to-view control group. The members of that group, as indeed preschoolers with access to public television throughout the nation were viewing the show with considerable frequency—and not much less than the experimental group. Thus, the planned analysis involving differences in posttest attainments between the two groups was dealt a mortal blow.

Fortunately, other analyses were available, of which the ETS-refined Age Cohorts Design provided a rational basis. This design is presented in the relevant report (Ball and Bogatz, 1970). The need here is not to describe the new design and analysis but to emphasize a point made practically by Robert Burns some time ago and repeated here more prosaically: The best laid plans of evaluators can gang agley, too.

Clearing new paths. Sometimes program evaluators find that the design and analysis they have in mind represent an untrodden path. This is perhaps in part because many of the designs in the social sciences are built upon laboratory conditions and simply are not particularly relevant to what happens in educational institutions.

When ETS designed the summative evaluation of *The Electric Company*, it was able to set up a true experiment in the schools. Pairs of comparable classrooms within a school and within a grade were designated as the pool with which to work. One of each pair of classes was randomly assigned to view the series. Pretest scores were used as covariates on posttest scores, and in 1973 the first-year evaluation analysis was successfully carried out (Ball and Bogatz). The evaluation was continued through a second year, however, and as is usual in schools, the classes did not remain intact.

From an initial 200 classes, the children had scattered through many more classrooms. Virtually none of the classes with subject children contained only experimental or only control children from the previous year. Donald B. Rubin, an ETS statistician, consulted with a variety of authorities and found that the design and analysis problem for the second year of the evaluation had not been addressed in previous work. To summarize the solution decided on, the new pool of classes was reassigned randomly to *E* (experimental) or *C* (control) conditions so that over the two years the design was portrayable as:



Note: For Year II, EE represents children who were in E classrooms in Year I and again in Year II. That is, the first letter refers to status in Year I and the second to status in Year II.

Further, the pretest scores of Year II were usable as new covariates when analyzing the results of the Year II posttest scores.

Tailoring to the task. Unfortunately for those who prefer routine procedures, it has been shown across a wide range of ETS program evaluations that each design and analysis must be tailored to the occasion. Thus, Gary Marco, as part of the 1972 statewide educational assessment in Michigan, evaluated ESEA Title I program performance. He assessed the amount of exposure students had to various clusters of Title I programs, and he included control schools in the analysis. He found that a regression analysis model involving a correction for measurement error was an innovative move that best fit his complex configuration of data.

Garlie Forehand, Marjorie Ragosta, and Donald A. Rock, in a large-scale, national, correlational study of desegregation completed in 1976, obtained data on school characteristics and on student outcomes. The purposes of the study included defining indicators of effective desegregation and discriminating between more and less effective school desegregation programs. The emphasis throughout the effort was on variables that were manipulable. That is, the idea was that evaluators would be able to suggest practical advice on what schools can do to achieve a productive desegregation program. Initial investigations allowed specification among the myriad variables of a hypothesized set of causal relationships, and the use of path analysis made possible estimation of the strength of hypothesized causal relationships. On the basis of the initial correlation matrices, the path analyses, and the observations made during the study, an important product—a nontechnical handbook for use in schools—was developed.

Another large-scale ETS evaluation effort was directed by Trismen, M.A. Waller, and Gita Wilder. They studied compensatory reading programs, initially surveying more than 700 schools across the country. Over a four-year period ending in 1976, this evaluation interspersed data analysis with new data collection efforts. One purpose was to find schools that provided exceptionally positive or negative program results. These schools were visited "blind" and observed by ETS staff. Whereas the Forehand evaluation analysis was geared to obtaining practical applications, the equally extensive evaluation analysis of Trismen's study was aimed at generating hypotheses to be tested in a series of smaller experiments.

As a further illustration of the complex interrelationship among evaluation purposes, design, analyses, and products, there is the 1977 evaluation of the use of PLATO in the elementary school by Spencer Swinton and Marianne Amarel. They used a form of regression analysis—as did Forehand and Trismen. But here the regression analyses were used differently in order to identify program effects unconfounded by teacher differences. In this regression analysis, teachers became fixed effects, and contrasts were fitted for each within-teacher pair (experimental versus control classroom teachers).

This, in turn, provides a contrast to McDonald's 1977 evaluation of West New York programs to teach English as a second language to adults. In this instance, the regression analysis was directed toward showing which teaching method related most to gains in adult students' performance.

There is a school of thought within the evaluation profession that design and analysis in program evaluation can be made routine. At this point, the experience of ETS indicates that this would be unwise.

Interpreting the Results

Possibly the most important principle in program evaluation is that interpretations of the evaluation's meaning—the conclusions to be drawn—are often open to various nuances. Another problem is that the evidence on which the interpretations are based may be inconsistent. The initial premise of this article was that the role of program evaluation is to provide evidence for decision-makers. Thus, one could argue that differences in interpretation, and inconsistencies in the evidence, are simply problems for the decision-maker and not for the evaluator.

But consider, for example, an evaluation by Powers in 1974 and 1975 of a year-round program in a school district in Virginia. (The long vacation was staggered around the year so that schools remained open in the summer.) The evidence presented by Powers indicated that the year-round school program provided a better utilization of physical plant and that student performance was not negatively affected. The school board considered this evidence as well as other conflicting evidence provided by Powers that the parents' attitudes were decidedly negative. The board made up its mind, and (not surprisingly) scotched the program. Clearly, however, the decision was not up to Powers. His role was to collect the evidence and present it systematically.

Keeping the process open. In general, the ETS response to conflicting evidence or varieties of nuances in interpretation is to keep the evaluation process and its reporting as open as possible. In this way, the values of the evaluator, though necessarily present, are less likely to be a predominating influence on subsequent action.

Program evaluators do, at times, have the opportunity to influence decision-makers by showing them that there are kinds of evidence not typically considered. The Coleman Study, for example, showed at least some decision-makers that there is more to evaluating school programs than counting (or calculating) the numbers of books in libraries, the amount of classroom space per student, the student-teacher ratio, and the availability of audiovisual equipment. Rather, the output of the schools in terms of student performance was shown to be generally superior as evidence of school program performance.

Through their work, evaluators are also able to educate decision-makers to consider the important principle that educational treatments may have positive effects for some students and negative effects for others—that an interaction of treatment with student should be looked for. As pointed out in the discussion of unintended outcomes, a system-analysis approach to program evaluation—dealing empirically with the interrelatedness of all the factors that may affect performance—is to be preferred. And this approach, as Messick emphasizes, "properly takes into account those student-process-environment interactions that produce differential results" (p. 246).

Selecting appropriate evidence. Finally, a consideration of the kinds of evidence and interpretations to be provided decision-makers leads inexorably to the realization that different kinds of evidence are needed, depending on the decision-maker's problems and the availability of resources. The most "scientific" evidence involving objective data on student performance can be brilliantly interpreted by an evaluator, but it might also be an abomination to a decision-maker who really needs to know whether teachers' attitudes are favorable.

ETS, over the past 10 years, has provided a great variety of evidence. For a formative evaluation in Brevard County, Florida, in 1970, Trismen provided evidence that students could make intelligent choices about courses. In the ungraded schools, students had considerable freedom of choice, but they and their counselors needed considerably more evidence than in traditional schools about the ingredients for success in each of the available courses. In 1977, Gary Echternacht helped state and local education authorities develop Title I reporting models that included evidence on impact, cost, and compliance with federal regulations. Forehand and McDonald have been working with New York City to develop an accountability model providing constructive kinds of evidence for the city's school system. On the other hand, as part of an evaluation team, Amarel is providing, for a small experimental school in Chicago, judgmental data as well as reports and documents based on the school's own records and files. And Michael Rosenfeld, in 1973, provided Montgomery Township, New Jersey, with student, teacher, and parent perceptions in his evaluation of the open classroom approach then being tried out.

In short, just as tests are not valid or invalid (it is the ways tests are used that deserve such descriptions), so, too, evidence is not good or bad until it is seen in relation to the purpose for which it is to be used, and in relation to its utility to decision-makers.

Postscript

For the most part, ETS's involvement in program evaluation has been at the practical level. Without an accompanying concern for the theoretical and professional issues, however, practical involvement would be irresponsible. ETS has seen the need to integrate and systematize its growing knowledge about program evaluation. Thus, Anderson obtained a contract with the Office of Naval Research to draw together the accumulated knowledge of professionals from inside and outside ETS to the topic of program evaluation. A number of products followed. These included a survey of practices in program evaluation and a codification of program evaluation principles and issues. Perhaps the most generally useful of the products is the aforementioned *Encyclopedia of Educational Evaluation*.

From an uncoordinated, nonprescient beginning in the mid-1960s, ETS has acquired a great deal of experience in program evaluation. In one sense it remains uncoordinated because there is no specific "party line," no dogma designed to ensure ritualized responses. It remains quite possible for different program evaluators at ETS to recommend differently designed evaluations for the same burgeoning or existing programs.

There is no sure knowledge where the profession of program evaluation is going. Perhaps, with zero-based budgeting, program evaluation will experience amazing growth over the next decade, growth that will dwarf its current status (which already dwarfs its status of a decade ago). Or perhaps there will be a revulsion against the use of social scientific techniques within the political, value-dominated arena of program development and justification. At ETS, the consensus is that continued growth is the more likely event. And with the staff's variegated backgrounds and accumulating expertise, ETS hopes to continue making significant contributions to this emerging profession.

Appendix A

Descriptions of ETS Studies in Some Key Categories

Aesthetics and Creativity in Education

For Bartlett Hayes III's program of Education through Vision at Andover Academy, Donald A. Trismen developed a battery of evaluation instruments that assessed, inter alia, a variety of aesthetic judgments. Other ETS staff members working in this area have included Norman Frederiksen and William C. Ward, who have developed a variety of assessment techniques for tapping creativity and scientific creativity; Richard T. Murphy, who also has developed creativity-assessing techniques; and Scarvia B. Anderson, who described, in 1968, a variety of ways to assess the effectiveness of aesthetic displays.

Bilingual Education

ETS staff have conducted and assisted in evaluations of numerous and varied programs of bilingual education. For example, Berkeley office staff have evaluated programs in Calexico (Reginald A. Corder, Jr.), Hacienda-La Puente (Patricia Elias, Patricia Wheeler), and El Monte (Corder, S. Johnson). For the Los Angeles office, J. Richard Harsh evaluated a bilingual program in Azusa, and Ivor Thomas evaluated one in Fountain Valley. Donald E. Hood of the Austin office evaluated the Dallas Bilingual Multicultural Program. These evaluations were variously formative and summative, and covered bilingual programs that, in combination, served students from preschool (Fountain Valley) through 12th grade (Calexico).

Camping Programs

Those in charge of a school camping program in New York City felt that it was having unusual and positive effects on the students, especially in terms of motivation. ETS was asked to—and did—evaluate this program, using an innovative design and measurement procedures developed by Raymond G. Wasdyke and Jerilee Grandy.

Career Education

In this decade of heavy federal emphasis on career education, ETS has been—and is—involved in the evaluation of numerous programs in that field. For instance, Raymond G. Wasdyke helped the Newark, Delaware, school system determine whether its career education goals and programs were properly meshed. In Dallas, Donald Hood of the ETS regional staff assisted in developing goal specifications and reviewing evaluation test items for the Skyline Project, a performance contract calling for the training of high school students in 12 career clusters. Norman E. Freeberg developed a test battery to be used in evaluating the Neighborhood Youth Corps. Ivor Thomas of the Los Angeles office provided formative evaluation services for the Azusa Unified School District's 10th grade career training and performance program for disadvantaged students. Roy Hardy of the Atlanta office directed the third-party evaluation of Florida's Comprehensive Program of Vocational Education for Career Development, and Wasdyke evaluated the Maryland Career Information System. Reginald A. Corder, Jr., of the Berkeley office assisted in the evaluation of the California Career Education program and subsequently directed the evaluation of the Experience-Based Career Education Models of a number of regional education laboratories.

Computer-aided Instruction

Three major computer-aided instruction programs developed for use in schools and colleges have been evaluated by ETS. The most ambitious is PLATO from the University of Illinois. Initially, the ETS evaluation was directed by Ernest Anastasio, but later the effort was divided between Richard T. Murphy, who focused on college-level programs in PLATO, and Spencer Swinton and Marianne Amarel, who focused on elementary and secondary school programs. ETS also directed the evaluation of TICCIT, an instructional program for junior colleges that uses small-computer technology; the study was conducted by Donald L. Alderman. Currently, Marjorie Regosta is directing the evaluation of the first major in-school longitudinal demonstration of computer-aided instruction for low-income students.

Drug Programs

Robert F. Boldt served as a consultant on the National Academy of Science's study assessing the effectiveness of drug antagonists (less harmful drugs that will "fight" the impact of illegal drugs). Samuel Ball served on a National Academy of Science panel that designed, for the National Institutes of Health, a means of evaluating media drug information programs and spot advertisements.

Educational Television

ETS was responsible for the national summative evaluation of the ETV series *Sesame Street,* for preschoolers, and *The Electric Company*, for reading students in grades one through four; the principal evaluators were Samuel Ball, Gerry Ann Bogatz, and Donald B. Rubin. Additionally, Ronald Flaugher and Joan Knapp evaluated the series *Bread and Butterflies* to clarify career choice; Jayjia Hsia evaluated a series on the teaching of English for high school students and a series on parenting for adults.

Higher Education

Much ETS research in higher education focuses on evaluating students or teachers, rather than programs, mirroring the fact that systematic program evaluation is not common at this level. ETS has made, however, at least two major forays in program evaluation in higher education. In their Open University study, Rodney T. Hartnett and associates joined with three American universities (Houston, Maryland, and Rutgers) to see if the British Open University's methods and materials were appropriate for American institutions. Mary Jo Clark, Leonard L. Baird, and Hartnett conducted a study of means of assessing quality in doctoral programs. They established an array of criteria for use in obtaining more precise descriptions and evaluations of doctoral programs than the prevailing technique—reputational surveys—provides.

Preschol Programs

A number of preschool programs have been evaluated by ETS staff, including the ETV series *Sesame Street*. Irving Sigel conducted formative studies of developmental curriculum. Virginia Shipman helped the Bell Telephone Companies evaluate their day care centers, and Samuel Ball and Brent Bridgeman provided the U.S. Office of Child Development with a sophisticated design for the evaluation of Parent-Child Development Centers.

Prison Programs

In New Jersey, ETS has been involved in the evaluation of educational programs for prisoners. Developed and administered by Mercer County Community College, the programs have been subject to ongoing study by Ronald L. Flaugher and Samuel Barnett.

Reading Programs

ETS evaluators have been involved in a variety of ways in a variety of programs and proposed programs in reading. For example, in an extensive, large-scale, national evaluation completed in 1976, Donald A. Trismen studied the effectiveness of reading instruction in compensatory programs. At the same time, Donald E. Powers, conducted a small study of the impact of a local reading program in Trenton, New Jersey. Ann M. Bussis, Edward A. Chittenden, and Marianne Amarel, in 1976, reported the results of their study of primary school teachers' perceptions of their own teaching behavior. Earlier, Richard T. Murphy surveyed the reading competencies and needs of the adult population.

Special Education

Samuel Ball and Karla Goldman conducted an evaluation of the largest private school for the learning disabled in New York City, and Carol Vale of the ETS office in Berkeley directed a national needs assessment concerning educational technology and special education. Paul Campbell is directing a major study of an intervention program for learning disabled juvenile delinquents.

Appendix B

References

Alderman, D. L. *Evaluation of the TICCIT computer-assisted instructional system in the community college*. Princeton, NJ: Educational Testing Service, 1978.

Amarel, M., and the Evaluation Collective. *Reform, response, renegotiation: Transitions in a school-change project*. Manuscript submitted to the Ford Foundation, 1979.

Anastasio, E. J. *Evaluation of the PLATO and TICCIT computer-based instructional systems—a preliminary plan*. PR-72-19. Princeton, NJ: Educational Testing Service, 1972.

Anderson, S. B. Noseprints on the glass—or how do we evaluate museum programs? In E. Larrabee (Ed.). *Museums and education*. Washington, DC: Smithsonian Institution Press, 1968, pp. 115-126.

Anderson, S. B. From textbooks to reality: Social researchers face the facts of life in the world of the disadvantaged. In J. Hellmuth (Ed.). *Disadvantaged child. Vol. 3: Compensatory education: A national debate*. New York: Brunner/Mazel, 1970.

Anderson, S. B., Ball, S., & Murphy, R. T. (Eds.). *Encyclopedia of educational evaluation*. San Francisco, CA: Jossey-Bass Publishers, 1975.

Ball, S. (contributor). *Evaluation drug information programs—report of the panel on the impact of information on drug use and misuse, phase 2*. Washington, DC: National Research Council, National Academy of Sciences, July 1973.

Ball, S., & Anderson, S. B. *Practices in program evaluation: A survey and some case studies*. Princeton, NJ: Educational Testing Service, October 1975 (a).

Ball, S., & Anderson, S. B. *Professional issues in the evaluation of education/training programs*. Princeton, NJ: Educational Testing Service, October 1975 (b).

Ball, S., & Bogatz, G. A. *The first year of Sesame Street: An evaluation*. PR-70-15. Princeton, NJ: Educational Testing Service, October 1970.

Ball, S., & Bogatz, G. A. *Reading with television: An evaluation of The Electric Company*. PR-73-2. Princeton, NJ: Educational Testing Service, February 1973.

Ball, S., Bogatz, G. A., Kazarow, K. M., & Rubin, D. B. *Reading with television: A follow-up evaluation of The Electric Company*. PR-74-15. Princeton, NJ: Educational Testing Service, June 1974.

Ball, S., Bridgeman, B., & Beaton, A. E. *A design for the evaluation of the parent-child development center replication project*. Princeton, NJ: Educational Testing Service, March 1976.

Ball, S., & Goldman, K. S. *The Adams School: An interim report*. Princeton, NJ: Educational Testing Service, February 1976.

Ball, S., & Kazarow, K. M. *Evaluation of To Reach a Child*. Princeton, NJ: Educational Testing Service, 1974.

Bogatz, G. A. Field operations. In S. B. Anderson, S. Ball, and R. T. Murphy (Eds.). *Encyclopedia of educational evaluation*. San Francisco, CA: Jossey-Bass Publishers. 1975. pp. 169-175.

Bogatz, G. A. & Ball, S. *The second year of Sesame Street: A continuing evaluation*. PR-71-21. Princeton, NJ: Educational Testing Service, November 1971.

Boldt, R. F (with N. Gitomer). Editing and scaling of instrument packets for the clinical evaluation of narcotic antagonists. PR-75-12. Princeton, NJ: 1975.

Bussis, A. M., Chittenden, E. A., & Amarel, M. *Beyond surface curriculum. An interview study of teachers' understandings*. Boulder, CO: Westview Press. 1976.

Campbell, P. B. *Psychoeducational diagnostic services for learning disabled youths*. Proposal submitted to Creighton Institute for Business Law and Social Research. Princeton, NJ: Educational Testing Service, December 1976.

Clark, M. J., Hartnett, R. Y., & Baird, L. L. *Assessing dimensions of quality in doctoral education*. PR-76-27. Princeton, NJ: Educational Testing Service, October 1976.

Corder, R. A., & Johnson, S. *Final evaluation report, 1971-1972, MANO A MANO*. Berkeley, CA: Educational Testing Service, January 1972.

Corder, R. A. Final evaluation report of part C of the California career education program. Berkeley, CA: Educational Testing Service, January 1975.

Corder, R. A. External evaluator's final report on the experience-based career education program. Berkeley, CA: Educational Testing Service, January 1976.

Corder, R. A. Calexico intercultural design. El cid: Title VII yearly final evaluation reports for grades 7-12 of program of bilingual education, 1970-1976. Berkeley, CA: Educational Testing Service, 1970-1976.

Dyer, H. S. A plan for evaluating the quality of educational programs in Pennsylvania. Harrisburg, PA: State Board of Education, 1965, Vol. 1, pp. 1-4; pp. 10-12; and Vol. 2, pp. 158-161.

Echternacht, G. (Princeton), Temp, G. (Atlanta), & Storlie, T. (Evanston). The operation of an ESEA Title I evaluation technical assistance center—Region 2. Proposal submitted to DHEW/OE, July 1976. Final report due in 1978.

Ekstrom, R. B., French, J., & Harman, H. Kit of reference tests for cognitive factors. 1976 edition. Princeton, NJ: Educational Testing Service, 1976.

Esktrom, R. B., & Lockheed, M. Giving women college credit where credit is due. Princeton, NJ: Findings, Vol. 3, No. 3, 1976.

Elias, P., & Wheeler, P. Interim evaluation report: BUENO. Berkeley, CA: Educational Testing Service, 1972.

Feldmesser, R. A. Educational goal indicators for New Jersey. PR-73-1. Princeton, NJ: Educational Testing Service, February 1973.

Flaugher, R. L. Progress report on the activities of ETS for the postal academy program. Unpublished manuscript. Princeton, NJ: Educational Testing Service, January 1971.

Flaugher, R., & Barnett, S. An evaluation of the Prison Educational Network. Unpublished manuscript. Princeton, NJ: Educational Testing Service, June 1972.

Flaugher, R, & Knapp, J. Report on evaluation activities of the Bread and Butterflies project. Princeton, NJ: Educational Testing Service, 1974.

Forehand, G. A., & McDonald, F. J. A design for an accountability system for the New York City school system. Princeton, NJ: Educational Testing Service, June 1972.

Forehand, G. A., Ragosta, M., & Rock, D. A. Final report: Conditions and processes of effective school desegregation. PR-76-23. Princeton, NJ: Educational Testing Service, July 1976.

Freeberg, N. E. Assessment of disadvantaged adolescents: A different approach to research and evaluation measures. Journal of Educational Psychology. 1970, 61, 229-240.

Hardy, R. A. CIRCO: The development of a Spanish language test battery for preschool children. Paper presented at the Florida Educational Research Association, Tampa, 1975.

Hardy, R. Evaluation strategy for developmental projects in career education. Tallahassee, FL. Division of Vocational, Technical, and Adult Education. Florida Department of Education. June 1977.

Harsh, J. R. A bilingual/bicultural project. Azusa Unified School District evaluation summary. Los Angeles, CA: Educational Testing Service, 1975.

Hartnett, R. T., Clark, M. J., Feldmesser, R. A., Gieber, M. L., & Soss, N. M. The British Open University in the United States. Princeton, NJ: Educational Testing Service, June 1974.

Harvey, P. R. National College of Education bilingual teacher education project. Evanston, IL: Educational Testing Service. October 1974.

Holland, P. W., Jamison, D. T., & Ragosta, M. Project report No. 1—phase 1 final report research design. Princeton, NJ: Educational Testing Service, 1976.

Hood, D. E. Final audit report: Skyline career development center. Austin, TX: Educational Testing Service, August 1972.

Hood, D. E. Final audit report of the ESEA IV supplementary reading programs of the Dallas Independent School District. Bilingual education program, pp. 115-143. Austin, TX: Educational Testing Service, November 11, 1974.

Hsia, J. Proposed formative evaluation of a WNET/13 pilot television program: The Speech Class. Proposal submitted to Educational Broadcasting Corporation. February 1976.

Marco, G. L. Impact of Michigan 1970-71 grade 3 Title I reading programs. PR-72-5. Princeton, NJ: Educational Testing Service, March 1972.

McDonald, F. J., & Elias, P. Beginning teacher evaluation study, phase 2. The effects of teaching performance on pupil learning. Vol. 1. PR-76-6A. Princeton, NJ: Educational Testing Service, 1976.

McDonald, F. J. The effects of classroom interaction patterns and student characteristics on the acquisition of proficiency in English as a second language. PR-77-5. Princeton, NJ: Educational Testing Service, 1977.

Messick, S. The criterion problem in the evaluation of instruction: Assessing possible, not just intended outcomes. In M. Wittrock & D. Wiley (Eds.), The evaluation of instruction: Issues and problems. New York: Holt, Rinehart and Winston, 1970.

Messick, S. Medical model of evaluation. In S. B. Anderson, S. Ball, & R. T. Murphy (Eds.). Encyclopedia of educational evaluation. San Francisco, CA: Jossey-Bass Publishers, 1975, pp. 245-247.

Murphy, R. T. *Investigation of a creativity dimension.* RB-73-12. Princeton, NJ: Educational Testing Service, February 1973.

Murphy, R. T. *Adult functional reading study*. PR-73-48. Princeton, NJ: Educational Testing Service, 1973.

Murphy, R. T. *Evaluation of the PLATO 4 computer-based education system: Community college component*. Princeton, NJ: Educational Testing Service, 1977.

Pike, L. W. *Relationships among Peruvian, Chilean, and Japanese students' scores over a wide range of measures of English as a foreign language*. Unpublished manuscript. Princeton, NJ: Educational Testing Service, 1977.

Powers, D. E. *An approach of the New Approach Method.* PR-73-47. Princeton, NJ: Educational Testing Service, November 1973.

Powers, D. E. *The Virginia Beach extended school year program and its effects on student achievement and attitudes—first year report.* PR-74-25. Princeton, NJ: Educational Testing Service, October 1974.

Powers, D. E. *The second year of year-round education in Virginia Beach: A follow-up evaluation*. PR-75-27. Princeton, NJ: Educational Testing Service, December 1975.

Powers, D. E. *Dual audio television: An evaluation of a six-month public broadcast.* PR-75-21. Princeton, NJ: Educational Testing Service, October 1975.

Rosenfeld, M. *An evaluation of the Orchard Road School open space program.* PR-73-14. Princeton, NJ: Educational Testing Service, June 1973.

Shipman, V. C. *Disadvantaged children and their first school experiences*. Vol. 1. PR-70-20. Princeton, NJ: Educational Testing Service, August 1970.

Shipman, V. C. *Evaluation of an industry-sponsored child care center. An internal ETS report prepared for Bell Telephone Laboratories.* Murray Hill, NJ. July 1974.

Sigel, I. E. *Developing representational competence in preschool children: A preschool educational program. In Basic needs, special needs: Implications for kindergarten programs. Selected papers from the New England Kindergarten Conference, Boston, December 5, 1975.* Cambridge, MA: The Lesley College Graduate School of Education. 1976.

Swinton, S., & Amarel, M. *The PLATO elementary demonstration: Educational outcome evaluation.* PR-78-11. Princeton, NJ: Educational Testing Service, 1978.

Thomas, I. J. *A bilingual and bicultural model early childhood education program. Fountain Valley School District Title VII bilingual project.* Berkeley, CA: Educational Testing Service, 1970.

Thomas, I. J. Mathematics aid for disadvantaged students. Los Angeles, CA: Educational Testing Service, 1973.

Trismen, D. A. *Evaluation of the Education through Vision curriculum—phase 1*. Princeton, NJ: Educational Testing Service, March 1968.

Trismen, D. A. (with T. A. Barrows). *Brevard County project: Final report to the Brevard County (Florida) school system.* PR-70-6. Princeton, NJ: Educational Testing Service, May 1970.

Trismen, D. A., Waller, M. A., & Wilder, G. *A descriptive and analytic study of compensatory reading programs.* Vols. 1 & 2. PR-76-3. Princeton, NJ: Educational Testing Service, February 1976.

Vale, C. A. *National needs assessment of educational media and materials for the handicapped.* Proposal submitted to Office of Education. Princeton, NJ: Educational Testing Service, September 1975.

Ward, W. C., & Frederiksen, N. *Development of measures for the study of creativity.* RB-75-18. Princeton, NJ: Educational Testing Service, May 1975.

Ward, W. C., & Frederiksen, N. *A study of the predictive validity of the tests of scientific thinking*. RB-77-6. Princeton, NJ: Educational Testing Service, June 1977.

Wasdyke, R. G., & Grandy, J. *Field evaluation of Manhattan Community School District #2 environmental education program.* Princeton, NJ: Educational Testing Service, December 1976.

Wasdyke, R. G. Year 3—*Third party annual evaluation report: Career education instructional system project.* Newark School District. Newark, Delaware. Princeton, NJ: Educational Testing Service, March 1977.

Wasdyke, R. G. *An evaluation of the Maryland Career Information System*. Oral report: August 1976.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. *Unobtrusive measures: Nonreactive research in the social sciences.* Chicago, IL: Rand McNally. 1966.

Woodford, P. E. *Pilot project for oral proficiency interview tests of bilingual teachers and tentative determination of language proficiency criteria.* Proposal submitted to Illinois State Department of Education. Princeton, NJ: Educational Testing Service, March 1975.