Program Evaluation Particularly Responsive Evaluation*

Originally published as Paper #5, Occasional Paper Series, November 1975.

* Paper presented at a conference on "New Trends in Evaluation", Goteborg, Sweden, October, 1973.

Robert E. Stake
Center for Instructional Research and Curriculum Evaluation
University of Illinois at Urbana-Champaign

I am pleased to have this opportunity to talk about some recent developments in the methodology of program evaluation and about what I call "responsive evaluation."

I feel fortunate to have not only these two days but also some seven months to think about these things. My hosts here at the G6teborg Institute of Educational Research have been most hospitable, but generous also in hearing me out, pointing my head in still another way, weighing the merit of our several notions, and offering occasionally the luxury of a passionate argument.

When Erik or Hans or Sverker or Ulf and I agree, we are struck by the fact that the world is but one world and the problems of education are universal. When we disagree, they are quick to suggest that the peculiar conditions of education in America have caused me to make peculiar assumptions and perhaps even warped my powers of reasoning. I am sure that some of you here today will share those findings. What I have to say is not only that we in educational research need to be doing some things we have not been doing, but that in doing what we have been doing we are in fact part of the problem.

Our main attention will be on program evaluation. A program may be strictly or loosely defined. It might be as large as all the teachers training in the United States or it might be as small as a field trip for the pupils of one classroom. The evaluation circumstances will be these: that someone is commissioned in some way to evaluate a program, probably an ongoing program; that he has some clients or audiences to be of assistance to—usually including the educators responsible for the program; and that he has the responsibility for preparing communications with these audiences.

In 1965 Lee Cronbach, then president of the American Educational Research Association, asked me to chair a committee to prepare a set of standards

for evaluation studies, perhaps like the Standards for Educational and Psychological Tests and Manuals, compiled by John French and Bill Michael and published in 1966 by the American Psychological Association. Lee Cronbach, Bob Heath, Tom Hastings, Hulda Grobman, and other educational researchers have worked with many of the U. S. curriculum-reform projects in the 1950's and early 1960's, and have recognized the difficulty of evaluating curricula and the great need for guidance on the design of evaluation studies.

Our committee reported that it was too early to decide upon a particular method or set of criteria for evaluating educational programs, that what educational researchers needed was a period of field work and discussion to gain more experience in how evaluative studies could be done. Ben Bloom, successor to Lee Cronbach in the presidency of AERA, got the AERPI to sponsor a Monograph Series on Curriculum Evaluation for the purpose we recommended. The seven volumes completed under AERA sponsorship are shown in the Reference section. The series in effect will continue under sponsorship of the UCLA Center for the Study of Evaluation, whose director, Mary Alkin, was a guest professor here at this Institute for Educational Research two years ago. I think this Monograph Series can take a good share of the credit, or blame, for the fact that by count over two hundred sessions at the 1973 AERA Annual Meeting programs were directly related to the methods and results of program-evaluation studies.

There are two primary models for program evaluation in 1965, and there are two today. One is the informal study, perhaps a self-study, usually using information already available, relying on the insights of professional persons and respected authorities. It is the approach of regional accrediting associations for secondary schools and colleges in the United States and is exemplified by the Flexner report (1916) of medical education in the USA and by the Coleman report (1966) of equality of educational opportunity. On the sheet you received with your background reading materials, one entitled Nine Approaches to Educational Evaluation (see Appendix A), I have ever so briefly described this and other models; this one is referred to there as the Institutional Self-Study by Staff Approach. Most educators are partial to this evaluation model, more so if they can specify who the panel members or examiners are. Researchers do not like it because it relies so much on secondhand information. But there is much good about the model.

Most researchers have preferred the other model, the pretest/posttest model, what I have referred to on the Nine Approaches sheet as Student Gain by Testing Approach. It often uses prespecified statements of behavioral objectives—such as are available from Jim Popham's Instructional Objectives Exchange—and is nicely represented by Tyler's (1942) Eight-Year Study, Husen's (1967) International Study of Achievement in Mathematics, and the National Assessment of Educational Progress. The focus of attention with this model is primarily on student performance.

Several of us have proposed other models. In a 1963 article is Cronbach's preference to have evaluation studies considered applied research on instruction, to learn what could be learned in general about curriculum development, as was done in Hilda Taba's Social Studies Curriculum Project. Mike Scriven (1967) strongly criticized Cronbach's choice in

AERA Monograph No. 1, stating that it was time to give consumers (purchasing agents, taxpayers, and parents) information on how good each existing curriculum is. To this end, Kenneth Komoski established in New York City an Educational Products Information Exchange, which has reviewed equipment, books, and teaching aids but has to this day still not caught the buyer's eye.

Dan Stufflebeam was one who recognized that the designs preferred by researchers did not focus on the variables that educational administrators have control over. With support from Egon Guba, Dave Clark, Bill Gephart, and others (1971), he proposed a model for evaluation that emphasized the particular decisions that a program manager will face. Data-gathering would include data on Context, Input, Process, and Product; but analyses would relate those things to the immediate management of the program. Though Mike Scriven criticized this design too, saying that it had too much bias toward the concerns and the values of the educational establishment, this Stufflebeam CIPP model was popular in the U. S. Office of Education for several years. Gradually, it fell into disfavor because it was a bad model but partly because managers were unable or unwilling to examine their own operations as part of the evaluation. Actually, no evaluation model could have succeeded. A major obstacle was a federal directive which said that no federal office could spend its funds to evaluate its own work, that that could only be done by an office higher up. Perhaps the best examples of evaluation reports following this approach are those done in the Pittsburgh schools by Mal Provus and Esther Kresh.

Before I describe the approach that I have been working on—which I hope will someday challenge the two major models—I will mention several relatively recent developments in the evaluation business.

It is recognized, particularly by Mike Scriven and Ernie House, that co-option is a problem, that the rewards to an evaluator for producing a favorable evaluation report often greatly outweigh the rewards for producing an unfavorable report. I do not know of any evaluators who falsify their reports, but I do know many who consciously or unconsciously choose to emphasize the objectives of the program staff and to concentrate on the issues and variables most likely to show where the program is successful. I often do this myself. Thus the matter of "meta-evaluation," providing a quality control for the evaluation activities, has become an increasing concern.

Early in his first term of office President Nixon created a modest Experimental Schools Program, a program of five-year funding for three carefully selected high schools (from all those in the whole country) and the elementary schools that feed students into them. Three more have been chosen each year, according to their proposal to take advantage of a broad array of knowledge and technical developments and to show how good a school can be. The evaluation responsibility was designed to be allocated at three separate levels, one internal at the local school level; one external at the local school level (i.e., in the community attending to the working of the local school but not controlled by it); and a third at the national level, synthesizing results from the local projects and evaluating the organization and effects of the Experimental Schools Program as a whole. Many obstacles and hostilities hampered the work of the first two evaluation teams. And work at the third level—according to Egon Guba, who did a

feasibility study—was seen to be so likely to fail that it probably should be carried no further.

Mike Scriven has made several suggestions for meta-evaluation, one most likely circulated based on abstinence, called "goal-free evaluation." Sixten Markiund has jokingly called it "aimless evaluation." But it is a serious notion, not to ignore all idea of goals with the program sponsors or staff. The evaluator, perhaps with the help of colleagues and consultants, then is expected to recognize manifest goals and accomplishments of the program as he works it in the field. Again, with the concern for the consumer of education, Scriven has argued that what is intended is not important, that the program is a failure if its results are so subtle that they do not penetrate the awareness of an alert evaluator. Personally I fault Scriven for expecting us evaluators to be as sensitive, rational, and alert as his designs for evaluation require. I sometimes think that Mike Scriven designs evaluation studies that perhaps only Mike Scriven is capable of carrying out.

Another interesting development is the use of adversarial procedures in obtaining evidence of program quality and especially in presenting it to decision makers. Tom Owens, Murray Levine, and Marilyn Kourilsky have taken the initiative here. They have drawn up the work of legal theorists who claim that truth emerges when opposing forces submit their evidence to cross-examination directly before the eyes of judges and juries. Graig Gjerde, Terry Denny, and I tried something like this in our TCITY report (1975). You have a copy of it in the conference reading materials you received several weeks ago. If you have that orange-colored document with you, you might turn to the very last pages, pages 26 and 27 (see Appendix B). On page 26 you find a summary of the most positive claims that might reasonably be made for the Institute we were evaluating. On page 27 is a summary of the most damaging charges that might reasonably be made. It was important to us to leave the issue unresolved, to let the reader decide which claim to accept, if any. But we would have served the reader better if we had each written a follow-up statement to challenge the other's claims. At any rate, this is an example of using an adversary technique in an evaluation study.

Now in the next 45 minutes or so I want to concentrate on the approach for evaluating educational programs presently advocated by Malcolm Parlett of the University of Edinburgh, Barry MacDonald of the University of East Anglia, Lou Smith of Washington University of St. Louis, Bob Rippey of the University of Connecticut, and myself. You have had an opportunity to read an excellent statement by Malcolm Parlett and David Hamilton (1972). Like they did, I want to emphasize the settings where learning occurs, teaching transactions, judgment data, holistic reporting, and giving assistance to educators. I should not suggest that they endorse all I will say today, but their writings for the most part are harmonious with mine.

Let me start with a basic definition, one that I got from Mike Scriven. Evaluation is an OBSERVED VALUE compared to some STANDARD. It is a simple ratio, but this numerator is not simple. In program evaluation it pertains to the whole constellation of values held for the program. And the denominator is not simple for it pertains to the complex of expectations and criteria that different people have for such a program.

The basic task for an evaluator is made barely tolerable by the fact that he does not have to solve this equation in some numerical way nor to obtain a descriptive summary grade but needs merely to make a comprehensive statement of what the program is observed to be, with useful references to the satisfaction and dissatisfaction that appropriately selected people feel toward it. Any particular client may want more than this; but this satisfies the minimum concept, I think, of an evaluation study.

If you look carefully at the TCITY report, you will find no direct expression of this formula, but it is in fact the initial idea that guided us. The form of presentation we used was chosen to convey a message about the Twin City Institute to our readers in Minneapolis and St. Paul rather than to be a literal manifestation of our theory of evaluation.

Our theory of evaluation emphasizes the distinction between a preordinate approach and a responsive approach. In the recent past the major distinction being made by methodologists is that between what Scriven called formative and summative evaluation. He gave attention to the difference between developing and already-developed programs and implicitly to evaluation for a local audience of a program in a specific setting as contrasted to evaluation for many audiences of a potentially generalizable program. These are important distinctions, but I find it even more important to distinguish between preordinate evaluation studies and responsive evaluation studies.

I have made the point that there are many different ways to evaluation educational programs. No one way is the right way. Some highly recommended evaluation procedures do not yield a full description nor a view of the merit and shortcoming of the program being evaluated. Some procedures ignore the pervasive questions that should be raised whenever educational programs are evaluated:

> Do all students benefit or only a specific few?
> Does the program adapt to instructors with unusual qualifications?
> Are opportunities for aesthetic experience realized?

Some evaluation procedures are insensitive to the uniqueness of the local condition. Some are insensitive to the quality of the learning climate provided. Each way of evaluating leaves some things de-emphasized.

I prefer to work with evaluation designs that perform a service. I expect the evaluation study to be useful to specific persons. An Evaluation probably will not be useful if the evaluator does not know the interests and language of his audiences. During an evaluation study a substantial amount of time may be spent learning about the information needs of the person for whom the evaluation is being done. The evaluators should have a good sense of whom he is working for and their concerns.

Responsive Evaluation

To be of service and to emphasize evaluation issues that are important for each particular program, I recommend the responsive evaluation approach. It is an approach that sacrifices some precision in measurement, hopefully to increase the usefulness of the findings to persons in and around the program. Many evaluation plans are more "preordinate," emphasizing (1) statement of goals, (2) use of objective tests, (3) standards held

by program personnel, and (4) research-type reports. Responsive evaluation is less reliant on formal communication, more reliant on natural communication.

Responsive evaluation is an alternative, an old alternative. It is evaluation based on what people do naturally to evaluate things: they observe and react. The approach is not new. But it has been avoided in planning documents and institutional regulations because, I believe, it is subjective, poorly suited to formal contracts, and a little too likely to raise the more embarrassing questions. I think we can overcome the worst aspects of subjectivity, at least. Subjectivity can be reduced by replication and operational definition of ambiguous terms even while we are relying heavily on the insights of personal observation.

An educational evaluation is responsive evaluation (1) if it orients more directly to program activities than to program intents, (2) if it responds to audience requirements for information, and (3) if the different value perspectives of the people at hand are referred to in reporting the success and failure of the program. In these three separate ways an evaluation plan can be responsive.

To do a responsive evaluation, the evaluator of course does many things. He makes a plan of observations and negotiations. He arranges for various persons to observe the program. With their help he prepares for brief narratives, portrayals, product displays, graphs, etc. He finds out what is of value to his audiences. He gathers expressions of worth from various individuals whose points of view differ. Of course, he checks the quality of his records. He gets program personnel to react to the accuracy of his portrayals. He gets authority figures to react to the importance of various findings. He gets audience members to react to the relevance of his findings. He does much of this informally, iterating, and keeping a record of action and reaction. He chooses media accessible to his audiences to increase the likelihood and fidelity of communication. He might prepare a final written report; he might not—depending on what he and his clients have agreed on.

Purpose and Criteria

Many of you will agree that the book edited by E. F. Lindquist, Educational Measurement, has been the bible for us who have specialized in educational measurement. Published in 1950, it contained no materials on program evaluation. The second edition, edited by Bob Thorndike (1971), has a chapter on program evaluation. Unfortunately, the authors of this chapter, Alex Astin and Bob Panos, chose to emphasize but one of the many purposes of evaluation studies. They said that the principal purpose of evaluation is to produce information that can guide decisions concerning the adoption of modification of an educational program.

People expect evaluation to accomplish many different purposes:

to document events
to record student change
to detect institutional vitality
to place the blame for trouble
to aid administrative decision making
to facilitate corrective action
to increase our understanding of teaching and learning

Each of these purposes is related directly or indirectly to the values of a program and may be a legitimate purpose for a particular evaluation study. It is very important to realize that each purpose needs separate data; all the purposes

cannot be served with a single collection of data. Only a few questions can be given prime attention. We should not let Astin and Panos decide what questions to attend to, or Tyler, or Stake. Each evaluator, in each situation, has to decide what to attend to. The evaluator has to decide.

On what basis will he choose the prime questions? Will he rely on his preconceptions? Or on the formal plans and objectives of the program? Or on actual program activities? Or on the reactions of participants? It is at this choosing than an evaluator himself is tested.

Most evaluators can be faulted for over-reliance on preconceived notions of success. I advise the evaluator to give careful attention to the reasons the evaluation was commissioned, then to pay attention to what is happening in the program, then to choose the value questions and criteria. He should not fail to discover the best and worst of program happenings. He should not let a list of objectives or an early choice of data-gathering instruments draw attention away from the things that most concern the people involved.

Many of my fellow evaluators are committed to the idea that good education results in measurable outcomes: student performance, mastery, ability, and attitude. But I believe it is not always best to think of the instrumental value of education as a basis for evaluating it. The "payoff" may be diffuse, long delayed; or it may be ever beyond the scrutiny of evaluators. In art education, for example, it is sometimes the purpose of the program staff or parent to provide artistic experiences—and training—for the intrinsic value alone. "We do these things because they are good things to do," says a ballet teacher. Some science professors speak similarly about the experimental value of reconstructing certain classical experiments. The evaluator or his observers should note whether or not those learning experiences were well arranged. They should find out what appropriately selected people think are the "costs" and "benefits" of these experiences in the dance studio or biology laboratory. The evaluator should not presume that only measurable outcomes testify to the worth of the program.

Sometimes it will be important for the evaluator to do his best to measure student outcomes, other times not. I believe that there are few "critical" data in any study, just as there are few "critical" components in any learning experience. The learner is capable of using many pathways, many tasks, to gain his measure of skill and aesthetic "benefit." The evaluator can take different pathways to reveal program benefit. Tests and other data-gathering should not be seen as essential; neither should they be automatically ruled out. The choice of these instruments in responsive evaluation should be made as a result of observing the program in action and of discovering the purposes important to the various groups having an interest in the program.

Responsive evaluations require planning and structure; but they rely little on formal statements and abstract representations, e.g. flow charts, test scores. Statements of objectives, hypotheses, test batteries, teaching syllabi are, of course, given primary attention if they are primary components of the instructional program. Then they are treated not as the basis for the evaluation plan but as components of the instructional plan. These components are to be evaluated just as other components are. The proper amount of structure for

responsive evaluation depends on the program and persons involved.

Substantive Structure

Instead of objectives or hypotheses as "advanced organizers" for an evaluation study, I prefer issues. I think the word issues better reflects a sense of complexity, immediacy, and valuing. After getting acquainted with a program, partly by talking with students, parents, taxpayers, program sponsors and program staff, the evaluator acknowledges certain issues or problems or potential problems. These issues are a structure for continuing discussions with clients, staff, and audiences. These issues are a structure for the data-gathering plan. The systematic observations to be made, the interviews and tests to be given, if any, should be those that contribute to understanding or resolving the issues identified.

In evaluating TCITY, Craig Gjerde and I became aware of such issue-questions as:

> Is the admissions policy satisfactory?
> Are some teachers too "permissive"?
> Why do so few students stay for the afternoon?
> Is opportunity for training younger teachers well used?
> Is this Institute a "lighthouse" for regular school curriculum innovation?

The importance of such questions varies during the evaluation period. Issues that are identified early as being important tend to be given too much attention in a preordinate data plan, and issues identified toward the end are likely to be ignored. Responsive-evaluation procedures allow the evaluator to respond to emerging issues as well as to preconceived issues.

The evaluator usually needs more structure than a set of questions to help him decide "what data to gather." To help the evaluator conceptualizes his "shopping list," I once wrote a paper entitled "The Countenance of Educational Evaluation" (Stake, 1967). It contained the matrix, the thirteen information categories, shown in this presentation on the screen (see Figure 1). You may notice that my categories are not very different from those called for in the models of Dan Stufflebeam and Mal Provus.

For different evaluation purposes there will be different emphases on one side of the matrix or the other: descriptive data and judgmental data. And, similarly, there will be different emphases on antecedent, transaction, and outcome information. The "Countenance" article also emphasized the use of multiple and even contradicting sources of information.

It also pointed out the often ignored question about the match-up between intended instruction and observed instruction and the even more elusive question about the strength of the contingency of observed outcomes upon observed transactions under the particular conditions observed. I think these "Countenance" ideas continue to be good ones for planning the content of the evaluation study.

I like to think of all of these data as observations: intents, standards, judgments, and statements of rationale are observed data too. Maybe it was a mistake to label just the second column "observations." Thoreau said: "Could a greater miracle take place than for us to look through each other's eyes for an instant."

Human observers are the best instruments we have for many evaluation issues. Performance data and preference data can be psychometrically scaled when

objectively quantified data are called for. The important matter for the evaluator is to get his information in sufficient amount from numerous independent and credible sources so that it effectively represents the perceived status of the program however complex.

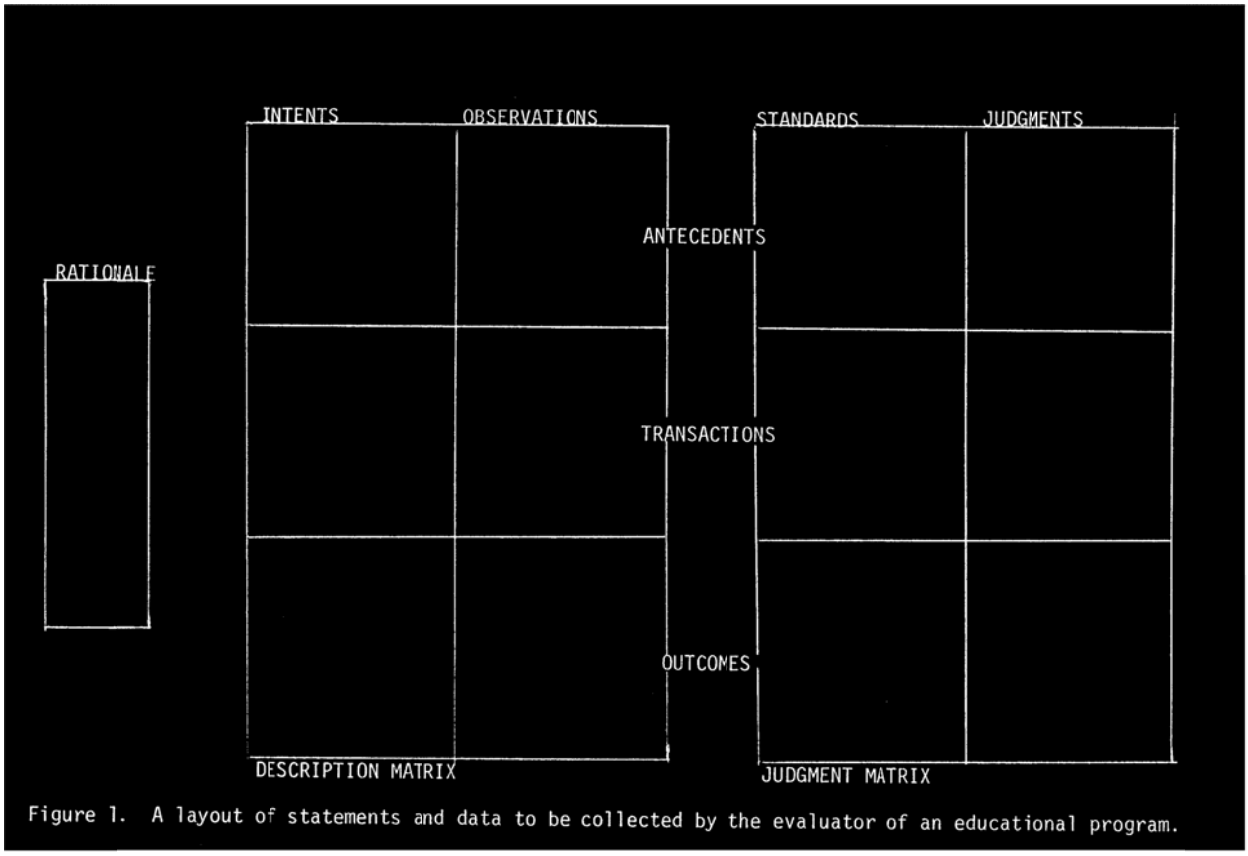

Figure 1. A layout of statements and data to be collected by the evaluator of an educational program.

Functional Structure

"Which data" is one thing but "how to do the evaluation" is another. My responsive-evaluation plan allocates a large expenditure of evaluation resources to observing the program. The plan is not divided into phases because observation and feedback continue to be the important functions from the first week through the last. I have identified twelve recurring events. On the screen here (see figure 2) I show them as if on the face of a clock. I know some of you would remind me that a clock moves clockwise so I hurry to say that this clock moves clockwise and counter-clockwise and cross-clockwise. In other words, any event can follow any event. Furthermore, many events occur simultaneously; and the evaluator returns to each event many times before the evaluation ends.

For example, take twelve o'clock. The evaluator will discuss many things on many occasions with the program staff and with people who are representative of his audiences. He will want to check his ideas of program scope, activities, purposes, and issues against theirs. He will want to show them his representations (e.g., sketches, displays, portrayals, photographs, tapes) of value

questions, activities, curricular content, and student products. Reactions to these representations will help him learn how to communicate in this setting. He should provide useful information. He should not pander to desires for only favorable (or only unfavorable) information nor should he suppose that only the concerns of evaluators and external authorities are worthy of discussion. (Of course, these admonitions are appropriate for responsive evaluation and preordinate evaluation alike.)

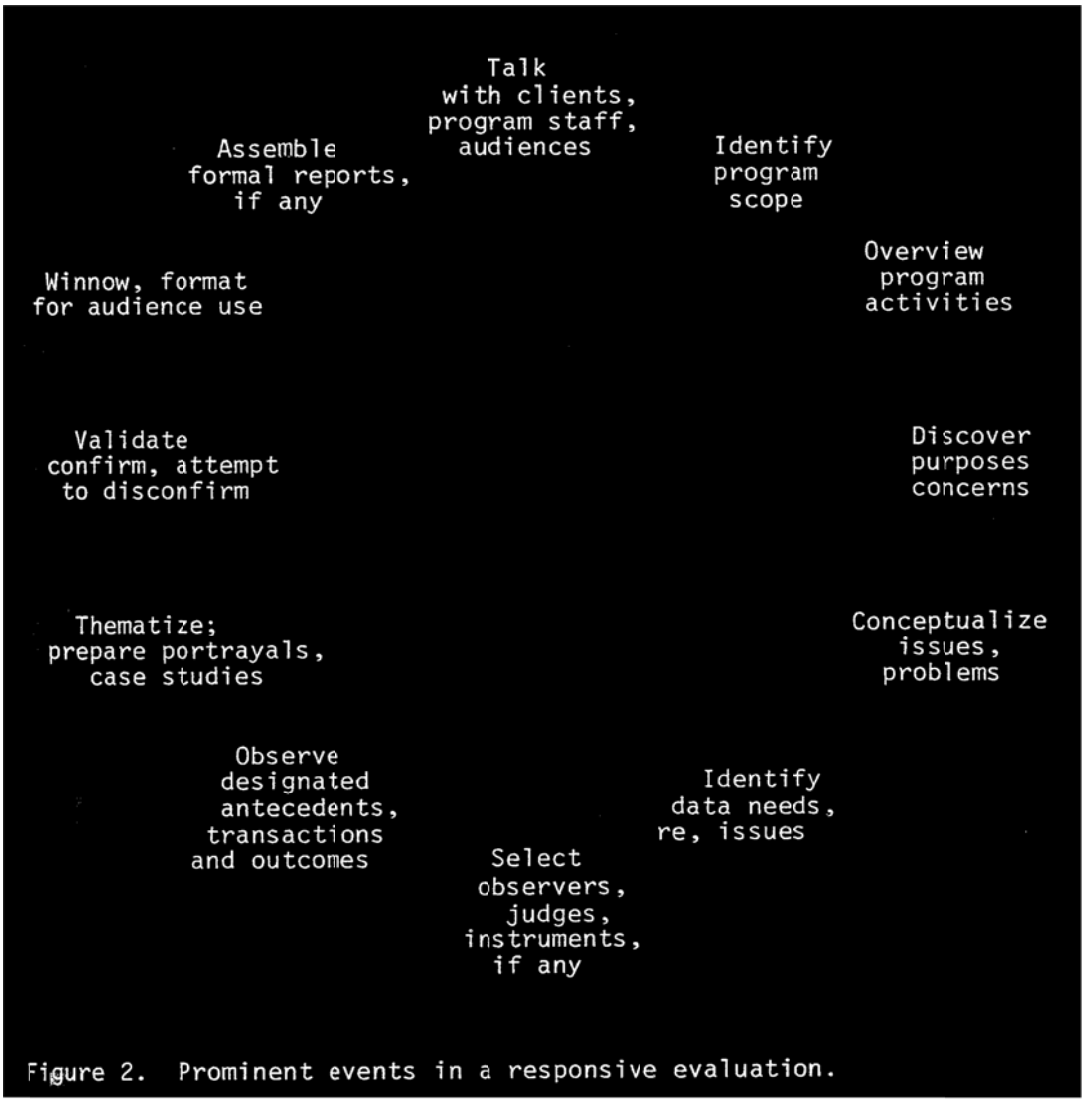

Figure 2. Prominent events in a responsive evaluation.

This behavior of the responsive evaluator is very different from the behavior of the preordinate evaluator. Here on the screen now (see below) is my estimate as to how the two evaluators would typically spend their time.

	Preordinate	Responsive
Identifying issues, goals	10%	10%
Preparing instruments	30%	15%
Observing the program	5%	30%
Administering tests, etc.	10%	—
Gathering judgments	—	15%
Learning client needs, etc.	—	5%
Processing formal data	25%	5%
Preparing informal reports	—	10%
Preparing formal reports	20%	10%

I believe the preordinate evaluator conceptualized himself as a stimulus, seldom as a response. He does his best to generate standardized stimuli, such as behavioral objective statements, test items, or questionnaire items. The responses that he evokes are what he collects as the substance of his evaluation report.

The responsive evaluator considers the principal stimuli to be those naturally occurring in the program, including responses of students and the subsequent dialogues. At first his job is to record these, learning both of happenings and values. For additional information, he assumes a more interventionist role. And, with his clients and audience he assumes a still more active role, stimulating their thought (we hope) and adding to their experience with his reports.

Philosopher David Hawkins (1973) responded to the idea of reversing S-R roles in this way:

> ...like the observation that one is reversing the S and R of it. In an experiment one puts the system in a prepared state, and then observes the behavior of it. Preparation is what psychologists call "stimulus,".. .In naturalistic investigation one does not prepare the system, but looks for patterns, structures, significant events, as they appear under conditions not controlled or modified by the investigator, who is himself now a system of interest. He is a resonator, a respondent. He must be in such an initial state that (a) his responses contain important information about the complex of stimuli he is responding to, and (b) they must be

> maximally decodable by his intended audience.

In the next section of this paper, I will talk about maximally decodable reports. Let me conclude these two sections on structure by saying that the evaluator should not rely only on his own powers of observation, judgment, and responding. He should enlist a platoon of students, teachers, community leaders, curriculum specialists, etc.—his choice depending on the issues to be studied and the audiences to be served. The importance of their information, and the reliability of it, will increase the number and variety of observers increase.

Portrayal and Holistic Communication

Maximally decodable reports require a technology of reporting that we educational measurements people have lacked. We have tried to be impersonal, theoretical, and generalizable. We have sought the parsimonious explanation. We have not accepted the responsibility for writing in a way that is maximally comprehensible to practicing educators and others concerned about education. According to R.F. Rhyne (1972):

> There is a great and growing need for the kind of powers of communication that helps a person gain, vicariously, a feeling for the natures of fields too extensive and diverse to be directly experienced.
> Prose and its archetype, the mathematical equation, do not suffice. They offer more specificity within a sharply limited region of discourse than is safe, since the clearly explicit can be so easily mistaken for truth, and the difference can be large when context is slighted (p. 93-104).

We need this power of communication, this opportunity for vicarious experience, in our attempts to solve educational problems.

One of the principal reasons for backing away from the preordinate approach to evaluation is to improve communication with audiences. The conventional style of research-reporting is a "clearly explicit" way of communicating. In a typical research project the report is limited by the project design. A small number of variables are identified and relationships among them are sought. Individuals are observed, found to differ, and distributions of scores are displayed. Covariations of various kinds are analyzed and interpreted. From a report of such analytic inquiry, it is very hard, often impossible, for a reader to know "what the program was like." If he is supposed to learn "what the program was like," the evaluation report should be different from the conventional research report.

As a part of my advocacy of the responsive approach I have urged my fellow evaluators to respond to what I believe are the natural ways in which people assimilate information and arrive at understanding. Direct personal experience is an efficient, comprehensive, and satisfying way of creating understanding but is a way not usually available to our evaluation report audiences. The best substitute for direct experience probably is vicarious experience—increasingly better when the evaluator uses "attending" and "conceptualizing" styles similar to those that members of the audience use. Such styles are not likely to be those of the specialist in measurement or the theoretically minded social scientist. Vicarious experience often will be conceptualized in terms of persons, places, and events.

We need a reporting procedure for facilitating vicarious experience. And it is available. Among the better evangelists, anthropologists, and dramatists are those who have developed the art of storytelling. We need to portray complexity. We need to convey holistic impression, the mood, even the mystery of the experience. The program staff or people in the community may be "uncertain." The audiences should feel that uncertainty. More ambiguity rather than less may be needed in our reports. Oversimplification obfuscates. Ianesco said (Esslin, 1966):

> As our knowledge becomes separated from life, our culture no longer contains ourselves (or only an insignificant part of ourselves) for it forms a "social" context into which we are not integrated.
> So the problem becomes that of bringing our life back into contact with our culture, making it a living culture once again. To achieve this, we shall first have to kill "the respect for what is written down in black and white..." to break up our language so that it can be put together again in order to re-establish contact with "the absolute," or as I should prefer to say, with "multiple reality"; it is imperative to "push human beings again towards seeing themselves as they really are" (P. 298).

Some evaluation reports should reveal the "multiple reality" of an educational experience.

The responsive evaluator will often use portrayals. Some will be short, featuring perhaps a five-minute "script," a log, or scrapbook. A longer portrayal may require several media: narratives, maps and graphs, exhibits, taped conversations, photographs, even audience role playing. Which ingredients best convey the sense of the program to a particular audience? The ingredients are determined by the structure chosen by the evaluator.

> Suppose that a junior-high-school art program is to be evaluated. For portrayal of at least one issue, "how the program affects every student," the students might be thought of as being in two groups: those taking at least one fine-arts course and those taking none. (The purpose here is description, not comparison.)
> A random sample of ten students from each group might be selected and twenty small case studies developed. The prose description of what each does in classes of various kinds (including involvement with the arts in school) might be supplemented with such things as (1) excerpts from taped interviews with the youngster, his friends, his teachers, and his parents; (2) art products (or photographs, news clippings, etc., of same) made by him in or out of class; (3) charts of his use of leisure time; and (4) test scores of his attitudes toward the arts. A display (for each student) might be set up in the gymnasium which could be examined reasonably thoroughly in 10-20 minutes.
> Other materials, including the plan, program, and staffing for the school, could be provided. Careful attention would be directed toward finding out how the description of these individual youngsters reveals what the school and other sources of art experience are providing in the way of art education.

It will sometimes be the case that reporting on the quality of education will require a "two-stage" communication. Some audiences will not be able to take part in such a vicarious experience as that arranged in the example above. A surrogate audience may be selected. The evaluator will present his portrayals to them; then he will question them about the apparent activity, accomplishments, issues, strengths, and shortcomings of the program. He will report their reactions, along with a more conventional description of the program, to the true audiences.

> These twenty displays could be examined by people specially invited to review and

> respond to them. The reviewers might be students, teachers, art curriculum specialists, and patrons of the arts. They might also visit regular school activities, but most attention would be to the displays. These reviewers should be asked to answer such questions as "Based on these case studies, is the school doing its share of providing good quality art experience for all the young people?" and "Is there too much emphasis on disciplined creative performance and not enough on sharing the arts in ways that suit each student's own tastes?" Their response to these portrayals and questions would be a major part of the evaluation report.

The portrayal will usually feature descriptions of persons. The evaluator will find that case studies of several students may more interestingly and faithfully represent the educational program than a few measurements on all of the students. The promise of gain is two-fold: the readers will comprehend the total program, and some of the important complexity of the program will be preserved. The several students usually cannot be considered a satisfactory representation of the many—a sampling error is present. The protests about the sampling error will be loud; but the size of the error may be small, and it will often be a satisfactory price to pay for the improvement in communication.

There will continue to be many research inquiries needing social survey technology and exact specification of objectives. The work of John Tukey, Torsten Husen, Ralph Tyler, Ben Bloom, and James Popham will continue to serve as a model for such studies.

Often the best strategy will be to select achievement tests, performance tests, or observation checklists to provide evidence that prespecified goals were or were not achieved. The investigator should remember that such a preordinate approach depends of a capability to discern the accomplishment of those purposes, and those capabilities sometimes are not at our command. The preordinate approach usually is not sensitive to ongoing changes in program purpose, nor to unique ways in which students benefit from contact with teachers and other learners, or to dissimilar viewpoints that people have as to what is good and bad.

Eliot Eisner (1969) nicely summarized these insensitivities in AERA Monograph No. 3. He advocated consideration of expressive objectives—toward outcomes that are idiosyncratic for each learner and that are conceptualized and evaluated after the instructional experience; after a product, an awareness, or a feeling has become manifest, at a time when the teacher and learner can reflect upon what has occurred. Eisner implied that sometimes it would be preferable to evaluate the quality of the opportunity to learn—the "intrinsic" merit of the experience rather than the more elusive "payoff," to use Scriven's terms.

In my own writing on evaluation I have been influenced by Eisner and Scriven and others who have been dissatisfied with contemporary testing. We see too little good measurement of complex achievements, development of personal styles and sensitivities. I have argued that few, if any, specific learning steps are truly essential for subsequent success in any life's endeavors; I have argued that students, teachers, and other purposively selected observers exercise the most relevant critical judgments, whether or not their criteria are in any way explicit. I have argued also that the alleviation of instructional problems is most likely to be accomplished by the people most directly experiencing the problem, with aid and comfort perhaps

(but not with specific solutions or replacement programs) from consultants or external authorities. I use these arguments as assumptions for what I call the responsive evaluation approach.

Utility and Legitimacy

The task of evaluating an educational program might be said to be impossible if it were necessary to express verbally its purposes or accomplishments. Fortunately, it is not necessary to be explicit about aim, scope, or probable cause in order to indicate worth. Explication will usually make the evaluation more useful; but it also increases the danger of misstatement of aim, scope, and probable cause.

To layman and professional alike, evaluation means that someone will report on the program's merits and shortcomings. The evaluator reports that a program is "coherent," "stimulating," "parochial," and "costly." These descriptive terms are also value-judgment terms. An evaluation has occurred. The validity of these judgments may be strong or weak; their utility may be great or little. But the evaluation was not at all dependent on a careful specification of the program's goals, activities, or accomplishments. In planning and carrying out an evaluation study, the evaluator must decide how far to go beyond the bare bones ingredients: values and standards. Many times he will want to examine goals. Many times he will want to provide a portrayal from which audiences may form their own value judgments.

The purposes of the audiences are all-important. What would they like to be able to do with the evaluation of the program? Chances are they do not have any plans for using it. They may doubt that the evaluation study will be of use to them. But charts and products and narratives and portrayals do not affect people. With these devices, persons become better aware of the program, develop a feeling for its vital forces, a sense of its disappointments and potential troubles. They may be better prepared to act on issues such as a change of enrollment or a reallocation of resources. They may be better able to protect the program.

Different styles of evaluation will serve different purposes. A highly subjective evaluation may be useful but not be seen as legitimate. Highly specific language, behavioral tasks, and performance scores are considered by some to be more legitimate. In American, however, there is seldom a greater legitimacy than the endorsement of large numbers of audience-significant people. The evaluator may need to discover what legitimacies his audiences (and their audiences) honor. Responsive evaluation includes such inquiry.

Responsive evaluation will be particularly useful during formative evaluation when the staff needs help in monitoring the program, when no one is sure what problems will arise. It will be particularly useful in summative evaluation when audiences want an understanding of a program's activities, its strengths and shortcomings, and when the evaluator feels that is his responsibility to provide a vicarious experience.

Preordinate evaluation should be preferred to responsive evaluation when it is important to know if certain goals have been reached, if certain promises have been kept, and when predetermined hypotheses or issues are to be investigated. With greater focus and opportunity for preparation, preordinate

measurements made can be expected to be more objective and reliable.

It is wrong to suppose that either a strict preordinate design or responsive design can be fixed upon an educational program to evaluate it. As the program moves in unique and unexpected ways, the evaluation efforts should be adapted to them, drawing from stability and prior experience where possible, stretching to new issues and challenges as needed.

References

Coleman, James S. et al. Equality of Educational Opportunity Washington, D.C.: U.S. Department of Health, Education and Welfare, Office of Education, 1966

Cronbach, Lee. "Course Improvement Through Evaluation." Teachers College Record, 1963, 64, 672-683.

DuBois, Philip H. & Mayo, D. Douglas (Eds.). Research Strategies for Evaluating Training (AERA Monograph Series on Curriculum Evaluation). Chicago: Rand McNally & Co., 1970.

Eisner, Eliot W. "Instructional and Expressive Educational Objectives: Their Formulation and Use in Curriculum." In W. James Popham, Eliot W. Eisner, Howard J. Sullivan, & Louise Tyler, Instructional Objectives (AERA Monograph Series on Curriculum Evaluation). Chicago: Rand McNally & Co. , 1969

Esslin, Martin. The Theater of the Absurd. London: Eyre & Spotteswoode, 1966.

Flexner, Abraham. Medical Education in the United States and Canada. A report to the Carnegie Foundation for the Advancement of Teaching. New York: The Carnegie Foundation, 1910. Reprinted New York: Arno Press, 1970

Gallagher, James; Nuthall, Graham A., & Rosenshine, Barak. Classroom Observation (AERI4 Monograph Series on Curriculum Evaluation). Chicago: Rand McNally & Co., 1970

Grobman, Hulda. Evaluation Activities of Curriculum Projects (AERA Monograph Series on Curriculum Evaluation) Chicago: Rand McNally & Co., 1968.

Hawkins, David. University of Colorado, Boulder, Colorado, (Personal communications).

Husen, Torsten (Ed.) International Study of Achievement in Mathematics. New York: Wiley, 1967

Levine, Murray. "Scientific Method and the Adversary Model: Some Preliminary Suggestions," Evaluation Comment, 1973,4, 1-3

Lindquist, Everett, F. (Ed.). Educational Measurement. Washington: American Council on Education, 1950.

Lindvall, C.M. & Cox, Richard. Evaluation as a Tool in Curriculum Development (AERA Monograph Series on Curriculum Evaluation). Chicago: Rand McNally & Co., 1970

MacDonald, Barry, "Evaluation and the Control of Education." In D. Tawney (Ed.). Evaluation: The State of the Art. London: Schools Council, 1975.

Parlett, Malcomb, & Hamilton, David. Evaluation as Illumination: A New Approach to the Study of Innovatory Programs. Edinburgh: Centre for Research in the Educational Sciences, University of Edinburgh, Occasional Paper No. 9, 1972.

Provus, Malcolm. Discrepancy Evaluation. Berkeley, California: McCutchan, 1971.

Rhyne, R.F. "Communicating Holistic Insights," Fields Within Fields-— Within Fields, 1972, 5, 93—104.

Scriven, Michael S. “The Methodology of Evaluation.” In Ralph Tyler, Robert Gagne, & Michael Scriven (Eds.), Perspectives of Curriculum Evaluation (AERA Monograph Series on Curriculum Evaluation.) Chicago: Rand McNally & Co., 1967.

_____. “Goal-Free Evaluation.” In Ernest House (Ed). School Evaluation: The Politics and Process. Berkeley, California: McCutcffan, 1973.

Smith, Louis M. , & Pahland, Paul A. “Educational Technology and the Rural Highlands.” In Louis M. Smith, Four Examples: Economic, Anthropological Narrative, and Portrayal (AERA Monograph on Curriculum Evaluation). Chicago: Rand McNally & Co., 1974

Rippey, Robert M. (Ed.). Studies in Transactional Evaluation. Berkeley, California: McCutchan, 1973.

Stake, Robert E. “The Countenance of Educational Evaluation,” Teachers College Record, 68, 1967, 523- 540

_____, & Gjerde, Craig. An Evaluation of TCITY: The Twin City Institute for Talented Youth. Paper #l, Evaluation Report Series. Kalamazoo, Michigan: Evaluation Center, Western Michigan University, 1975.

Stufflebeam, Daniel L. et al. Educational Evaluation and Decision Making. Itasca, Illinois: Peacock, 1971

Thorndike, Robert L. (Ed.) Educational Measurement, Washington: American Council on Education, 1971.

Tyler, Ralph W. “Eight-Year Study.” In E. R. Smith and Ralph Tyler Appraising and Recording Student Progress. New York: Harper, 1942

_____. Basic Principles of Curriculum and Instruction.
Chicago: University of Chicago Press, 1949

Womer, Frank. What is National Assessment? Ann Arbor: National Assessment of Educational Progress, 1970

APPROACH	PURPOSE	KEY ELEMENTS	PURVIEW EMPHASIZED	PROTAGONISTS (see references)	CASES EXAMPLES	RISKS	PAYOFFS
STUDENT GAIN BY TESTING	to measure student performance and progress	goal statements; test scores analysis; discrepancy between goal and actuality	EDUCATIONAL PSYCHOL-OGISTS	Ralph Tyler Ben Bloom Jim Popham Mal Provus	STEELE WOMER LINDVALL-COX Husen	oversimplify, educ'1 aims; ignore processes	emphasize, ascertain student progress
INSTITUTIONAL SELF-STUDY BY STAFF	to review and increase staff effectiveness	committee work standards set by staff; discussion; professionalism	PROFESSORS TEACHERS	National Study of school Evaluation Dressel	BOERSMA-PLAWECKI KNOLL-BROWN CARPENTER	alienate some staff; ignore values of outsiders	increase staff awareness, sense of responsibility
BLUE-RIBBON PANEL	to resolve crises and preserve the institution	prestigious panel; the visit; review of existing and documents	LEADING CITIZENS	James Conant Clark Kerr David Henry	FLEXNER HAVINGHURST HOUSE ET AL PLOWDEN	postpone action; over-rely on intuition	gather best insights judgment
TRANSACTION-OBSERVATION	to provide understanding of activities and values	educational issues; classroom observation; case studies; pluralism	CLIENT, AUDIENCE	Lou Smith Parlett-Hamilton Rob Rippey Bob Stake	MacDONALD SMITH-POHLAND PARLETT LUNDGREN	over-reply on subjective perceptions; ignore causes	produce broad picture of program; see conflict in values
MANAGEMENT ANALYSIS	to increase rationality in day to day decisions	lists of options; estimates; feedback loops; costs; efficiency	MANAGERS, ECONOMISTS	Leon Lessinger Dan Stufflebeam Don Campbell	KRAFT DOUGHTY-STAKENAS HEMPHILL	over-value efficiency; undervalue implicits	feedback for decision making

INSTRUCTIONAL RESEARCH	to generate explanations and tactics of instruction	controlled conditions, multivariate analysis; bases for generalization	EXPERIMEN-TALISTS	Lee Cronbach Julian Stanley Don Campbell	ANDERSON, R. PELLA ZDEP-JOYCE TABA	Artificial conditions ignore the humanistic	new principles of teaching and materials development
SOCIAL POLICY ANALYSIS	to aid development of institutional policies	measures of social conditions and administrative implementation	SOCIOLOGISTS	James Coleman David Cohen Carol Weiss Monsteller-Moynihan	COLEMAN JENCKS LEVITAN TRANKELL	neglect of educational issues, details	social choices, constraints clarified
GOAL-FREE EVALUATION	to assess effects of program	ignore proponent claims, follow checklist	CONSUMERS	Michael Scriven	HOUSE-HOGBEN	over-value documents & record keeping	data on effect with little co-option
ADVERSARY EVALUATION	to resolve a two-option choice	opposing advocates, cross-examination, the jury	EXPERT, JURISTIC	Tom Owens Murray Levine Bob Wolfe	OWENS STAKE-GJERDE REINHARD	personalistic superficial, time-bound	info. impact good; claims put to test

Of course these descriptive tags are a great over simplification. The approaches overlap. Different proponents and different users have different styles. Each protagonist recognizes one approach is not ideal for all purposes. Any one study may include several approaches. The grid is an over-simplification. It is intended to show some typical, gross differences between contemporary evaluation activities.

Appendix B

TCITY-1971 Evaluation Report: An Advocate's Statement

No visitor who took a long, hard look at TCITY-7l kept his skepticism. A young visitor knows how precious it is to discover, to be heard, to belong. An older visitor knows the rarity of a classroom where teachers and students perceive each other as real people. To the non-visitor it doesn't seem possible that a summer school program can deliver on all these promises to over 800 kids, but TCITY-71 did.

Every curriculum specialist fears that by relaxing conduct rules and encouraging student independence they may be saying goodbye to the hard work and hard thinking that education requires. TCITY-71 teachers and students made learning so attractive, so purposive, that free-ranging thought returned again and again to curricular themes: awareness of the human condition, obstacles to communication, ecological interactions, etc.

TCITY excels because of its staff. Its students give it movement. Its directors give it nurture. Its teachers give it movement, nurture, and direction. It would be incorrect to say that Mr. Caruson, Mr. Rose, and the teachers think alike as to the prime goals and methods of education, but collectively, they create a dynamic, humanistically-bent, academically-based curriculum.

The quality of teaching this summer was consistently high, from day to day, from class to class. Some of the teachers chose to be casual, to offer "opportunities," to share a meaningful experience. Others were more intense, more intent upon sharing information and problem solving methods. Both kinds were there, doing it well.

The quality of the learning also was high. The students were tuned in. They were busy. They responded to the moves of their teachers. They improvised, they carried ideas and arguments, indignations and admirations, to the volleyball court, to the Commons, to the shade of campus elms and Cannon River oaks. The youngsters took a long step towards maturity.

True, it was a costly step. Thousands of hours, thousands of dollars, and at least a few hundred aggravations. But fit to a scale of public school budgets--and budgets for parks, interstate highways, and weapons of war--TCITY-7l rates as a BEST BUY. 800 kids, give or take a few, took home a new talent, a new line of thinking, a new awareness--a good purchase.

It cannot be denied that other youngsters in Minneapolis and St. Paul deserve an experience like this. They should have it. Some say, "TCITY is bad because it caters to the elite." But a greater wisdom says, "Any effort fixated on giving an equal share of good things to all groups is destined to share nothing of value." For less advantaged youth, a more equitable share of educational opportunities should be guaranteed. But even in times of economic recession, opportunities for the talented should be protected.

TCITY-71 has succeeded. It is even a best buy. It satisfies a social obligation to specially educate some of those who will lead-it, the arts, in business, in government, in life. The teachers of TCITY-71 have blended a summer of caring, caprice, openness, and intellectual struggle to give potential leaders a summer of challenge.

(Prepared by R. Stake, not to indicate his opinion of the Institute, but as a summary of the most positive claims that might reasonably be made.)

TCITY-197l Evaluation Report: An Adversary's Statement

TCITY is not a scandalum magnatum. But it is both less than it pretends to be and more than it wishes to be. There is enough evidence at least to question certain facets of the Institute--if not to return a true bill against it. Costly, enlarging, innovative, exemplary: these Institute attributes are worthy of critical examination.

How costly is this Institute? Dollar costs are sufficient to give each group of six students $1,000 to design and conduct their won summer experience. Over 100 Upward Bound students could be readied for their college careers at Macalester. About twenty-five expert curriculum specialists could be supported for half a year to design and develop new curricula for the high school.

What is the cost of removing 800 talented leaders from the local youth culture? What is the cost of widening the experience gap between Institute students and their parents? And their teachers in "regular" high school? And their non-Institute friends? Not enough here to charge neo-Facist elitism. Enough to warrant discussion.

The Institute abounds with self-named innovators and innovations, with alternatives to the business-as-usual education of high schoolers. Note that the Institute is not promoted as an exemplary alternative to schooling. It seeks to promote the development of alternative forms of education for schools. And it is failing to do even that job. What is TCITY doing to demonstrate that TCITY style of life could be lived in schools as we know them? Where in the regular school is the staff so crucial to the life of the Institute?... the money?. . .the administrative leadership? Where are the opportunities for the teachers, principals, superintendents to come and live that life that they might come to share in the vision?.. .and where are the parents? TCITY should be getting poor grades on affecting the regular school program.

There are other dimensions of TCITY that puzzle the non-believer:

> *** How long can in-class "rapping" continue and still qualify as educative self-exploration? Are there quality control procedures in effect during the summer program: For example: when one-third to one—half of a class is absent from a scheduled meeting, should not that be seen as an educational crisis by the instructor?
>
> *** What does TCITY do to help students realize that the Institute standards are necessarily high; that the regular schools norms and expectations do not count; that a heretofore "best" becomes just a "so-so"? There are unnecessarily disheartened students in TCITY.
>
> *** Is it unreasonable to expect that more than two of twenty-two teachers or associate teachers would have some clear idea or plan for utilizing TCITY approaches or curricula in their regular classrooms next fall?
>
> *** Few students--or faculty--understand the selection procedures employed to staff the teaching cadre and to fill the student corps. Why should it be a mystery?

The worst has been saved for last. This report concludes with an assertion: the absence of crucial dimension in the instructional life of TCITY, that of constructive self-criticism, is a near fatal flaw. The observation and interview notes

taken by the adversary evaluator over four days contain but five instances of students engaging in, or faculty helping students to become skillful in, or desirous of, the cultivation of self-criticism. The instances of missed opportunities were excessive in my judgment. Worse: when queried by the writer, faculty and students alike showed little enthusiasm for such fare. Is it too much to expect from Institute participants after but four weeks? Seven may be insufficient. The staff post mortem, "Gleanings," are a start--but it seems odd to start at the end.

The paucity of occurrence is less damning than the absence of manifest, widespread intent. Certain classes accounted for all the instances observed. They did not appear to be accidental. The intent was there. An Institute for talented high school youth cannot justifiably fail to feature individual and group self-criticism.

(Prepared by T. Denny, not to indicate his opinion of TCITY-1971, but as a summary of the most damaging charges that might reasonably be made.)