DOCUMENT RESUME

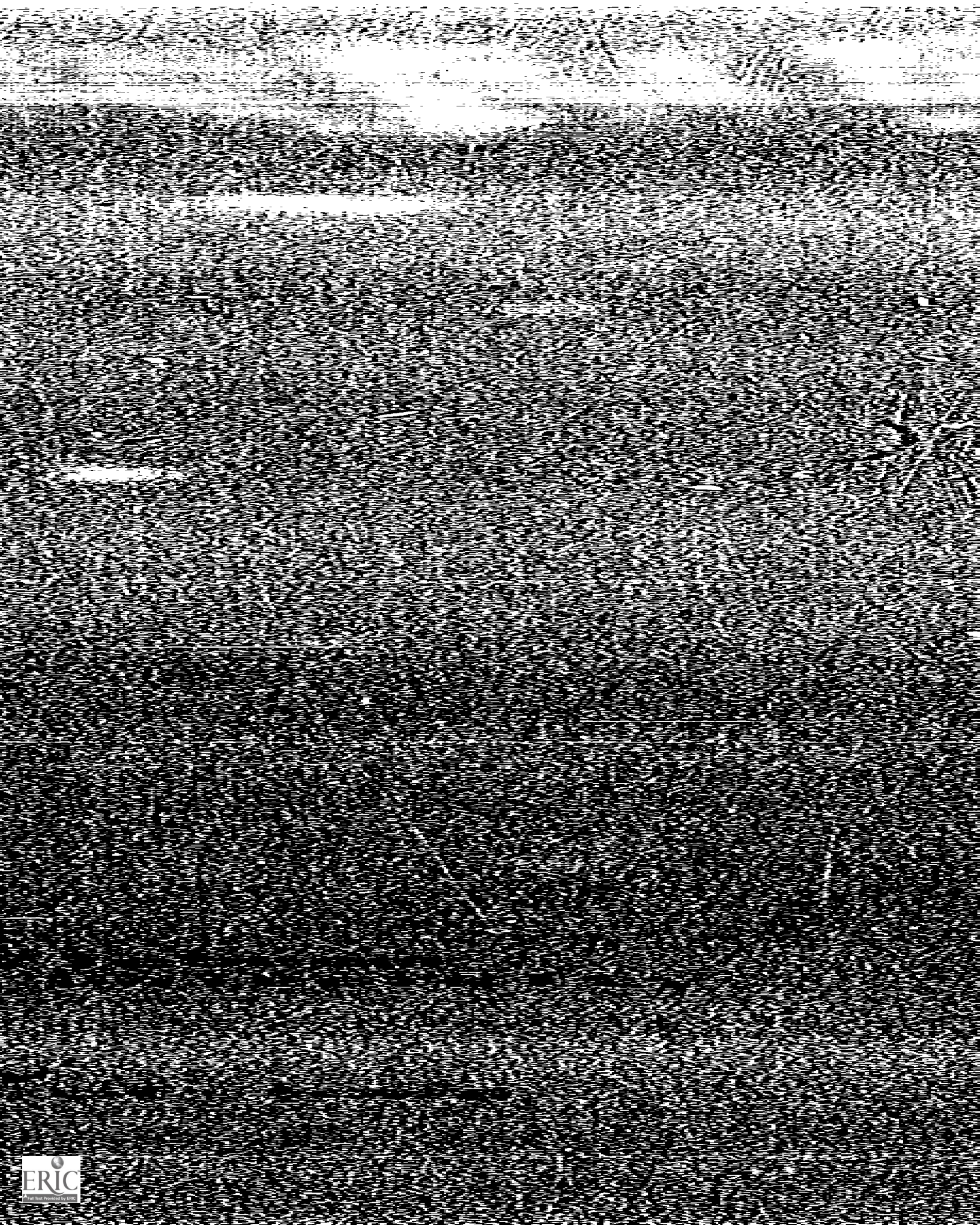
ED 056 074 TM 000 877

AUTHOR Stufflebeam, Daniel L.
TITLE Critique of the Report of the Phi Delta Kappa Study Committee on Evaluation.
PUB DATE 6 Feb 71
NOTE 59p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Book Reviews; *Decision Making; *Educational Accountability; Educational Planning; Evaluation; *Evaluation Methods; Evaluation Needs; Instrumentation; Measurement Techniques; Methodology; Models; *Symposia

ABSTRACT

An evaluative symposium conducted during the 1971 Annual Meeting of the AERA offered the following critiques on "Educational Evaluation and Decision Making," a book prepared by the Phi Delta Kappa Study Committee on Evaluation: "A Critique of the Report of the Phi Delta Kappa Study Committee on Evaluation" (Henry M. Brickell); "A Critique of the Measurement and Instrumentation Aspects of 'Educational Evaluation and Decision-Making' " (John C. Flanagan); "A Critique of the Methodology of Evaluation in the PDK Volume, 'Educational Evaluation and Decision Making' " (William B. Michael); "Evaluation: Noble Profession and Pedestrian Practice" (Michael Scriven); "Determining 'Most Probable' Causes: A Call for Re-examining Evaluation Methodology" (James L. Wardrop). (MS)



FOREWORD

Two symposia were conducted during the 1971 Annual Meeting of the American Educational Research Association to deal with a recent book entitled Educational Evaluation and Decision Making, prepared by the Phi Delta Kappa National Study Committee on Evaluation. The first symposium was descriptive in nature, while the second was evaluative. This report contains only the information from the second symposium, since the substance of the first symposium is already available through the Phi Delta Kappa book.

The first symposium was a description of a two-year effort by the PDK National Study Committee on Evaluation to analyze problems and to conceptualize relevant solutions in the field of evaluation. Members of the PDK Committee introduced and summarized the material contained in the eleven chapters of their final report. This symposium was intended to provide the basis for the second related symposium, in which experts in educational change theory, educational administration, educational psychology, philosophy of science, and educational evaluation offered critical reactions to the PDK report on evaluation.

The second symposium was chaired by Walter J. Foley, a member of the PDK Study Committee. The critiquers included Henry M. Brickell, Institute for Educational Development, John C. Flanagan, American Institutes for Research, William B. Michael, University of Southern California, Michael Scriven, University of California at Berkeley, and James L. Wardrop, University of Illinois at Urbana-Champaign.

Each critiquer had reviewed an advance copy of the PDK book and had developed a formal reaction. Copies of the critiques, as edited by the authors, form the substance of this report.

As organizer of the second symposium, I wish to express appreciation for the diligent efforts of the reviewers to provide in-depth reactions to the PDK report and for Walter Foley's capable chairing of the session. Important issues are identified among the critiques and should serve to further the efforts of those who are committed to improving both the theory and practice of educational evaluation.

DLS

A CRITIQUE OF THE REPORT OF THE PHI DELTA KAPPA STUDY COMMITTEE ON EVALUATION

Henry M. Brickell

The enormous scope of the Commission report is an achievement for its authors but a massive problem for its critics. Where does one grasp the beast to wrestle with it? For those of you in the audience who have not read the Commission report, the risk you face is that each one of us at the symposium will drag a different section into the ring, leaving the bulk of the creature outside the arena. That will give you the impression not that we are reporting on an elephant but that we have segmented a giant platypus and randomly assigned its parts for evaluation. I wanted you to have my assessment of the context we are in before hearing my input into the process of judging the product. (The Commission's report can change your language.)

The breadth of the Commission's work outreaches any critique of less than 532 pages, the length of the book itself. Its publication is an action that triggers reactions. It accuses; one wants to make counter accusations. It argues; one wants to argue back. It illustrates; one wants to use opposing illustrations. Where it asks a question, one wants to answer. When it gives an answer, one wants to question.

To set some limits around my own comments, I will view this new creature with the eyes of a practicing decision-maker in a public school district. There are two reasons for this. The first is that I have spent most of my career as a decision-maker and have had a good

chance to observe other decision-makers in action. The second reason is that there is a large and growing demand for evaluation services for those who manage either ongoing school programs or special projects and who must periodically decide whether to continue, modify, or terminate them. The obligation to evaluate ESEA Title I and other federal programs is of course a major source of that demand. But certain new developments, such as rising community interest in schools and the growth of performance contracting, increase the demand for evaluation services. Decision-makers, who are the clients for the kind of evaluation being proposed, will have definite opinions about the utility of those services. Since funds for the kind of evaluation being proposed by the Commission will come from those clients rather than from traditional research-funding sources, their reactions will shape the future of the movement. Research can be pursued at the initiative of the individual scholar, even without special funding, but the kind of evaluation envisioned by the Commission certainly cannot be. Thus, apart from the need for more detailed conception and a much better methodology (needs pointed to by the Commission itself) the reactions of decision-makers may be decisive.

Now I will react as a decision-maker. Allright. Here is a body of theory and practice that wants to be my servant--no, my consultant, or even more accurately, my colleague. My very first reaction is, "I have met someone like you before--in fact, quite a number of you." I have a school psychologist who often comes in when I am trying to make decisions and explains that I ought to use him as a consultant.

He tells me that the school is a wholly human enterprise, that he specializes in human behavior, and that in as much as my decisions deal with people he ought to guide me. The last time he was in I had to break off for an appointment with the curriculum coordinator, who explains that since instruction is the central function of the school, I ought to let him advise on all my major decisions. The day before, my business manager had reminded me that we run the place on money, that every decision I make has a price tag, and that he can keep us in the black only if I will bring him in on my decisions in advance. I didn't know whether to be more impressed with that or with what our community relations man had said about how he could keep us out of needless trouble with the activists if I would check with him beforehand on how my decisions would be received in the community. I had left the community relations specialists to meet with the building principals, who complain that I spend too much time listening to the central office staff. The real work of the system takes place out in the school buildings, they say, and they as principals have the best vantage point for helping me make decisions. The teachers' union has negotiated itself a chair on my side of the desk so that it can keep me acquainted in advance with how my decisions will be taken by teachers, which the union explains is only simple justice since the teachers are the people who must carry out decisions.

Once I thought I had command of every one's territory; now I realize that they all have command of mine.

But I must admit that you are especially intriguing to me, you new-breed evaluator. The immediate reason happens to be that I have just hired a planner. He has explained his job to me. I have learned from him that planning is the central decision-making function and that he will thus spend most of his time helping me make decisions. I am not surprised. But what interests me most is that he wants to help me 1) set goals, 2) choose among alternative courses of action he has generated, 3) allocate people and money to the chosen course of action, and 4) use the results to plan better next time. Obviously, not only are both of you in my territory, which is already crowded, you are also in each other's. So not only can I welcome you to the club, I even know who to give you for a roommate. Maybe the two of you can work out which one is going to help me do what.

Before I learned about your concept of evaluation, I had a fairly simple picture of how to use both a planner and an evaluator. I would send the planner out of one door of my office with a roll of plans under his arm and eventually you, as an evaluator, would come in the other door with an evaluation report on how the plans had worked out. You could meet the planner at my desk. But I realize now that he has stretched himself so far forward that he is standing at your door and you have stretched yourself so far backward that you are standing at his. I have had to deal with role conflict and now I can see I am going to have to deal with role overlap.

I have some other reactions as well. You remind me for some reason of the action research movement. As Max Corey described it,

action research in the hands of the classroom didn't sound to me much like research, but it did sound like intelligent action. For example if the teacher could tell that things were going wrong she was supposed to change them right then--not keep on going so she could accumulate a solid mass of dismal results at the .05 or preferably .01 level. That made sense; that's exactly the way I do things as a decision-maker. You do seem to understand that, unlike most of the evaluators I have met, I am not going to hold things steady just so you can evaluate them. But I have always felt the evaluators believed I was somewhat sloppy, that they couldn't help me if I wouldn't play the game their way, and that it was all my fault. If you could really help me without getting in the way and cramping my style while I am trying to run with the ball, then I'm interested. But I have enough other hurdles to jump over without having to clear some extra ones that you set up.

You remind me also for some reason of the new curriculum packages where the examination doesn't come at the end of the course but is scattered in pieces throughout, lesson by lesson or unit by unit so that the teacher can tell how things are going--even child by child--without waiting until it's too late. That makes sense. If you can do something like that about the decisions I have to make in my office, I'm interested.

One thing I may as well tell you quite frankly. I wouldn't even be considering you for a job if I didn't have these federal programs

that I have to evaluate. Your salary is going to come right out of that evaluation money and you are going to have to keep the state officials satisfied that our Title I projects are successful so that we can keep on getting the funds. I am willing to change projects that don't succeed, you understand, but what you have to know is that the state people are mainly interested in tested pupil performance--Washington pushes them that way, of course--and they don't want to settle for anything else. If your approach is going to add something to the evaluation of that ultimate product--pupil learning--but not try to substitute something for it, then I'm interested. But we have to keep the state people happy and I have to be sure they will settle for your method. Something I don't fully understand is how you are going to use the findings of the research done elsewhere, particularly the research that is generalizable to my situation. Are you saying that when I put in something that has been proven successful by previous research elsewhere that I still have to pay to have it evaluated all over again in my district? Why can't we use those other results?

The things you are talking about doing sound pretty expensive. But if I understand you correctly, we don't have to evaluate everything. At least, we don't have to put everything through the full evaluation cycle. We could evaluate only the major changes, we could evaluate all the changes. Or, if I could afford it,

we could evaluate everything whether it has been changed or not. That decision would be up to me and it would depend on what we could afford. Right?

One thing I am very concerned about is that you not go around turning up a lot of trouble. I think things are going pretty well. Anyhow I hope so--for several reasons. One is that the staff is working pretty hard under difficult conditions and I am not interested in having you bring in a critical report every couple of weeks finding fault with something. The teacher and principals need encouragement and a sense of success more than anything else. In addition to that, it would not make a very good impression on the Board if you found trouble in every corner. One big reason I hope you are not going to find many things wrong is that I can't keep on changing everything all the time. First the district can't afford it and second, I don't want the place in constant turmoil. So, if your approach can locate success as well as failure--and not show that nothing we try ever works, like most of our past evaluation consultants have found, then I am interested. At the very least, you will have to help me rank any problems you find so that I can solve the worst of them and let the others go by.

I have been paying a little attention to the performance contracting movement. We are not really interested right now and I want to see how it goes in other districts first, but I

would like to know whether your approach could put the finger on things that we would be better off contracting out to some other agency--or maybe even to our own teachers union. The Board of Education would certainly be interested in that.

One other thing you have to understand. I may not always take your advice, even if you think it's good. I have other things to think about. Let me give you an example. Last October an evaluation consultant I had hired brought in a report on our paraprofessionals for the previous year. He was able to show pretty convincingly that paraprofessionals weren't having any effect on the test scores of the elementary children, which we have been hoping they would. He said his findings forced him to recommend that the paraprofessional program be terminated in favor of something else. Great. Great advice. That's all he knew about it. What he didn't know was that we were in a bit of a recession in this city last fall. All I needed to do was to drop those minority-group paraprofessionals from the payroll--all of them live right around the schools where they work, have kids in school, and a lot of contacts in the neighborhood. Fire them all during that recession and we would have had something close to an armed revolution. So naturally, I still have the program going on just as before. I can't use advice like what he gave. Now, if there is something in the way you do evaluation that can take into account all aspects of the problem, then I'm interested.

So much for my reactions as a practicing decision-maker. Let me step out of his shoes and comment on the Commission's work as an evaluator, which I am from time to time. As an evaluator, I would welcome something that would enable me to stop saying to the decision-makers that come to me:

"Too late. You should have come to me long before you started this program. It's January, and you've been underway since September. Your achievement testing program doesn't contain anything I can use as a post-test, much less a pre-test. You didn't even specify pupil behavioral objectives when you started this International Exchange Program for Teachers. Moreover, it looks like everybody who could benefit from it is already in the program so we don't even have a control group. Sorry; I can't help you."

Well, I want to help. I don't want to be a sorry researcher if I can be a useful evaluator.

In summary, the work of the Commission represents a promising way of bringing disciplined inquiry into the service of the decision-maker, something researchers have had great difficulty in doing. The amount of intelligence and hard work applied by the Commission surely will advance us toward that objective.

Certainly we need something between the mindless and rapid evaluations performed in the early days of Title I and the excessive reactions to them which have government officials today expecting us to show that if a Teacher Corps Trainee goes to a good lecture

as a college junior in 1971, her pupils in 1973 will get higher achievement test scores as a result.

The Commission report itself is honest in trying to point out its own shortcomings. One of the best critiques is the self-examination contained in the final chapter of the volume.

What I admire most is the Commission's boldness in daring to call for work which requires the invention of new methodology--rather than inventing new work to fit the available methodology.

A CRITIQUE OF THE MEASUREMENT AND INSTRUMENTATION ASPECTS OF EDUCATIONAL EVALUATION AND DECISION-MAKING

John C. Flanagan

As one of a panel of five reactors to this repc I feel a ilttle bit like one of the five blind men describing the elep ant. Unlike the blind men, my colleagues and I all see this elephant, but our descriptions are likely to be rather dissimilar because of our special fields of interest and our varied previous experiences. Thus, the descriptions, although all based on the same 532-page 3-ton elephant, can be expected to be quite different.

In my description the emphasis will be on techniques of measurement, data collection, and the central role of the individual student in evaluation activities. Before proceeding to specific points, some general impressions seem in order. First, the report is comprehensive, detailed, and analytical. It analyzes evaluation into stages occurring in various settings, having various scopes, and providing information relevant to various types of decisions. It is thorough and systematic and provides a very good framework. Of course, it isn't the book I would have written because it doesn't really do much for providing measurement techniques, at least not new ones, or even a real good review of what we do have, and of course it doesn't center on the individual as the evaluation unit in the way that I would like to see it, although there is mention of this.

The report is based on a specific definition of evaluation which is: "Educational evaluation is the process of delineating, obtaining, and providing useful information for judging decision alternatives."

This definition is followed by a discussion of four stages in the process of decision-making including seventeen specific elements. In addition to this study of the decision process, there is a detailed description of possible decision settings, decision models, types of decisions, and some problems related to decision-making. In this chapter and the chapters which follow on criteria, values, information and systems theory, and evaluation methodology, the emphasis seems to be on delineating and discussing all possibilities rather than on the practical side of the conduct of educational evaluation.

In Chapter Seven, the four types of evaluation are presented together with a general model for conducting any one of these types of evaluation. The three steps which are proposed for all types of evaluation are delineating, obtaining, and providing.

The four types of evaluation are:

First, Context Evaluation which has as its purpose "to provide a rationale for determination of objectives. Specifically, it defines the relevant environment, describes the desired and actual conditions pertaining to that environment, identifies unmet needs and unused opportunities, and diagnoses the problems that prevent needs from

being met and opportunities from being used. The diagnosis problems provides an essential basis for developing objectives whose achievement will result in program improvement."

The authors state that context evaluation is the most basic kind of evaluation. The authors divide context evaluation into two modes - contingency and congruence. In the contingency mode evaluation searches for opportunities to improve the system by changing the objectives. The congruence mode evaluates the extent to which intended objectives are achieved.

This reviewer strongly endorses the emphasis on this type of evaluation and the distinction between the two modes for studying the objectives of the system. The discussion, however, seems to lack sufficient emphasis on needs and opportunities with respect to individual students and, although there is an emphasis on broad exploratory probing, it appears desirable that there be more specific provision for unplanned outcomes and the achievement of unintended objectives as well as those intended for the system.

The second type of evalution, input evaluation, is intended to provide the basis for selecting a design to achieve program objectives. This involves the study of relevant capabilities, strategies for achieving objectives, and basic specific designs for implementing a proposed strategy.

The authors point out that "techniques for input evaluation are lacking in education." One available technique which appears

applicable and is not discussed is the method of explicit rationales.

The third type of evaluation, process evaluation, is intended "to provide feedback to persons responsible for implementing plans and procedures." To a substantial degree, what these authors have included in process evaluation has come to be known as formative evaluation following the terminology of Michael Scriven. The objectives of process evaluation are: to monitor the implementation of the design, to provide information needed for planned decisions during the implementation phase, and to maintain a record of the extent to which the project is actually implemented as designed. This type of evaluation is clearly of great importance.

The fourth type of evaluation, product evaluation, measures and interprets the extent to which objectives were achieved. The criteria which are measured to perform this evaluation are classified as either instrumental or consequential following Scriven's terminology. Instrumental criteria refer to what have been frequently called intermediate criteria. Consequential criteria are those usually called ultimate criteria.

The authors point out that "in the assessment of objectives relating to adoption, product evaluation and context evaluation ultimately merge in the measurement of the impact of the total change effort on the overall system. Context evaluation then takes on the systematic functions of monitoring the total system and the ad hoc product evaluation is terminated."

In a later section the authors state "product evaluation assesses attainments of change projects within a system, and context evaluation assesses the impact of the obtained change on the total system." This distinction between context and product evaluation seems to be a useful one. In their general discussion of the features of their evaluation model, the authors again emphasize the basic importance of context evaluation and the need for a much more comprehensive data base to perform this function. Unfortunately, they do not seem to go far enough in developing specifications and procedures for collecting this very important data base. Educational systems have continued to operate with very little attention at either the local or national level to the study of the needs of individual students. The authors of this study have inserted two or three paragraphs suggesting the use of the individual student as the unit of measure in evaluation studies. The remarks are relevant and valuable. It would be desirable if their implications were carried through more fully in the subsequent discussions of implementing evaluation programs.

The later chapters of the report on implementing and administering evaluation programs need to be supplemented by handbook materials on what data to collect to study the needs and opportunities of the total educational system especially as it relates to the individual student. Some of the procedures used in recent years to obtain such

data include intensive case studies of students on a sampling basis; follow-up studies of recent graduates to determine the utility of the knowledge and abilities achieved in school; and intensive studies of adults in various roles and activities to determine the specific educational objectives which would have been most appropriate for them during their study programs in school. These types of data are not mentioned in the book.

There is probably no more important problem in education at the present time than determining the educational objectives for each individual student. It is believed that during the later educational years, much of the responsibility for these decisions should be given to the student. To prepare him for taking such responsibility it is believed that one should start in the primary grades by giving students some responsibility in planning and carrying out their educational programs. This will necessarily be limited in the early years, but the ability to take responsibility requires much practice.

This will require that the student know the specific knowledges and abilities required for many adult roles and activities. He must also know something about the nature of learning and individual differences and be able to estimate the extent of effort required for him to achieve a specific level of proficiency with respect to various types of content or ability. To assist the student

In formulating his long-range educational and occupational goals the behavioral scientist needs detailed and extensive studies of students both during and following their exposure to specific educational experiences which can be made available to current students as a basis for making their decisions. A minor point regarding the present report is that in view of these authors such behavioral scientists are clearly functioning as evaluators in providing the basis for individual decisions; however, their functions appear to be broadly those of the behavioral scientist and not specifically those usually considered as appropriate for an evaluator.

To sum up this review of educational evaluation and decision-making as presented by these authors, the first point to be noted is that the definition selected by these authors includes only one type of evaluation in education and therefore should not be thought of as the only function of evaluation methods in the educational field. There are many instances in which evaluative data are very desirable even though no decisions have been defined and no actions are anticipated. However, for purposes of decision oriented educational evaluation, the report has much to commend it. The efforts of the seven members of the Phi Delta Kappa Commission on Evaluation represent an important step forward in increasing our understanding and ability to conduct effective educational evaluation studies. As the authors point out, this is only the beginning of an important effort to improve our educational programs.

A CRITIQUE OF THE METHODOLOGY OF EVALUATION IN THE PDK VOLUME, EDUCATIONAL EVALUATION AND DECISION MAKING

William B. Michael

The PDK volume affords probably the most comprehensive and penetrating conceptualization of educational evaluation and decision making currently available. The CIPP (context-input-process-product) Model can use much of existing research methodology, particularly in product evaluation, but does require new methodology especially in reference to the context, input, and process components that tend to be somewhat more closely associated with formative evaluation than with product evaluation.

The modest stance which the PDK Evaluation Study Committee members group have taken and their receptivity to suggestions undoubtedly mean that there will be ample opportunity for the work group to make improvements and to move forward in developing the kinds of methodology that will be needed. They seem to be neither inflexible in their orientation to evaluation nor resistant to suggestions. For this kind of openmindedness they are to be commended.

One may look at the evaluation methodology relative to the CIPP Model from two standpoints: (1) the feasibility of the CIPP Model given current research methodology, and (2) the need for new methodology given the CIPP approach to evaluation. After consideration of each of these two broad topics, some recommendations will be set

forth which essentially will summarize many of the comments that are made in the critique of the first two categories.

Feasibility of the CIPP Model Given Current Methodology

Relative then to the first category, the feasibility of the CIPP Model given current research methodology, these comments may by offered:

1. In summative evaluation of products, as in accountability studies, classical research methodology involving use of an experimental and control group still affords a viable approach provided that evaluators can be placed in positions of influence and power in funding and governing agencies to allow them to require certain approved research designs--especially those designs incorporating use of randomization in large-scale evaluation studies. Such large-scale studies are expensive, but the educational enterprise is also expensive. Such carefully planned investigations carried out under relatively well controlled conditions would permit the formation of causal inferences and generalization as well as valid methods for determining cost-effectiveness in accountability studies.

2. Relative to existing research methodology, several other points may be noted.

a. A great deal of existing research methodology can be used in evaluation studies with quite specific and limited objectives as in the determination of the most effective ways to teach

multiplication or substraction or to acquire clearly delineated psychomotor skills.

b. Laboratory schools within colleges of education afford opportunities for experimentation in which products can be evaluated in a relatively reliable and valid manner.

c. Some research designs such as the multiple time series or regression discontinuity analysis described by Campbell and Stanley afford a basis for drawing causal inferences especially in relation to product-oriented evaluation and for making decisions regarding effectiveness of alternative educational programs.

d. In the laboratory orientation of many colleges of education opportunities exist to investigate systematically in an evaluation setting such important problems as differential rates of retention and learning, the transfer of learning which, too often, is glibly assumed to occur, and the effectiveness of group versus individual problem solving endeavor through use of simulation games.

Need for New Methodology Given the CIPP Model

Despite this rather positive state of affairs which allows one to use existing research methodology for product evaluation in a reasonably well controlled setting, there certainly is, however, a need for new methodology given the CIPP approach to evaluation. Especially in the context, input, and process forms of evaluation

which are often carried out in different social and environmental contexts, there is a definite need for new methodology, much of which might be adapted from anthropology, sociology, history, political science, and economics. Curriculum specialists and evaluators find the distinctions among these three components of context, input, and process evaluation to be somewhat overlapping and at times to be somewhat ambiguous or even contradictory.

The following points may be noted:

1. Throughout the recycling process of the CIPP Model, which involves constant feedback and dynamic modifications in the various steps of evaluation and decision-making, there is the need to establish either explicitly or implicity cause and effect relationships primarily through narrowing or limiting the number of possible alternative hypotheses. The work of Yee and Gage recently reported in the Psychological Bulletin offers considerable promise for establishing possible direction of cause and effect among several sets of correlated measures obtained at different times.

2. A pressing need exists to develop a methodology for establishing value systems in the selection of objectives and in their implementation particularly within the realm of context and input evaluation. After all, the word "evaluation" in essence contains the word "value." The problem of setting priorities in the selection

of objectives within a social context is particularly essential and deserving of serious systematic treatment. Again, methodological approaches underlying historical, anthropological, and sociological research could be potentially very helpful.

3. In all phases of the CIPP Model and throughout the whole process of decision making, consideration needs to be given to the ways in which different kinds of available information may be sorted, integrated, and incorporated in the decision-making process. Such information can be examined in a relatively well controlled simulation game or through observation in a realistic day-to-day school setting permitting replicability. Field tests need to be made of many of the procedures suggested in the CIPP Model, as Bob Hammond of Montana's Department of Public Instruction is doing just now. He is involving people with training in other kinds of disciplines. For example, he has one man with a background in banking who, in having a rather realistic orientation to cost accounting and cost-effectiveness, is trying out the CIPP Model in rather practical contexts. Furthermore, field testing will allow for extensive trial and error observation so that one will know how the various components of the CIPP Model will work.

4. Another underlying methodological concern is the need to distinguish between the role of the evaluator and the decision-maker and to ascertain in a given context for evaluation the relative degree of independence or overlap of their respective functions.

5. Once refinements in the methodology for use with the CIPP Model are developed, concerted efforts will be required to familiarize and train individuals in the use of the model. For example, participants in a two-day simulation game held at The Ohio State University in conjunction with an evaluation of the PDK volume last June said that they found the model was very useful as a conceptual framework but difficult to implement because of their lack of familiarity with it. They necessarily relied on past experience and intuition to make evaluation decisions. Thus, extensive study of and experience with the CIPP Model are needed so that one can use it rather automatically without having to ferret through all of the various components. Currently the complexities of the model will probably prevent its wide use and application until there is extensive in-service training.

It is important to point out that the heuristic properties of the CIPP Model for doctoral dissertation research are great indeed even if its use displaces the overworked classical experimental-control group models. Doctoral committees should encourage students to do developmental studies even though they may represent a break with tradition. As professors of educational research, many in the audience could encourage students to do developmental kinds of work such as that which could be used for validating the CIPP Model.

Recommendations

Within the context of what has been said the following recommendations may be formulated:

1. The CIPP Model should be given extensive consideration as a guide for conceptualizing essential characteristics of discrepancy evaluation activities.

2. Efforts should be made whenever possible to use existing research methodology in implementing many of the objectives of the CIPP Model, especially in the instance of product evaluation.

3. Attention should be directed toward devising new methodologies, many of which can be adapted from those of the social sciences, to answer questions raised by application of the CIPP Model. In particular, the following recommendations may be set forth:

a. Concerted effort should be followed throughout the stages of context, input, process, and product evaluation to furnish different kinds of evidence that will make possible formulation of both explicit and implicit inferences regarding possible cause and effect relationships.

b. Systematic efforts should be directed toward the development of a methodology for setting value systems in the selection and implementation of objectives as in context and input evaluation.

c. The full process of decision making in reference to the availability of many kinds of information should be studied systematically in relation to the CIPP evaluation model.

d. Attention should be given to how the CIPP Model can be advantageously used in accountability studies, for which there will be increasing pressures and demands.

4. Further energy needs to be directed to define and differentiate the educational and technical role of the evaluator, a distinction which has not been made entirely clear to the satisfaction of a number of persons who attended the PDK-The Ohio State University conference last June.

5. The feasibility of establishing seminars and in-service training institutes to give the CIPP Model greater visibility and utility to the school community should be investigated.

Summary Evaluation

All in all, the CIPP Model offers great promise of providing both external and internal validity of the evaluation process. Certainly the initial three steps of context, input, and process evaluation do much to sharpen the thinking of the evaluator who is oriented toward product evaluation, because the first three steps indeed afford the monitoring, recycling, and feedback functions upon which effective product evaluation depends. The external validity, however, is still open to serious concern, especially in the accountability studies. Threats to external validity may be due most often to a lack of randomization or to the inability of the evaluator to assume a position of power and influence which he

might assume in evaluation studies involving decisions about a multi-million dollar educational enterprise. Irrespective of the size of the evaluation effort or the magnitude of the decision-making processes involved, the CIPP Model probably affords the most comprehensive conceptualization of education currently available. The expenditure of efforts to develop new and to adopt existing methodologic for obtaining, analyzing, and interpreting the data which the model can generate should increase its usefulness in the education establishment.

EVALUATION: NOBLE PROFESSION AND PEDESTRIAN PRACTICE

Michael Scriven

The PDK report is certainly the most compendious and may well be the most valuable treatise on evaluation in the literature. Perhaps there should be a three-minute silence at this place, for that is almost the last nice thing I say. But that's just because it is not efficient to keep on saying nice things: the brain does not store redundant information.

I think there are significant flaws in its basic conception of evaluation as well as in its practical advice. Here are a few.

Its basic conception excludes crucial paradigms of evaluation: for example, the evaluation by historians of Napoleon's tactics at Waterloo. I take a non-educational case, for interest, but there can of course be educational ones too; suppose you are evaluating the school system in Athens. It seems clear that this is logically the same kind of enterprise as evaluating the tactics of an on-going field general or contemporary educational activities. But the latter can easily and the former can usually not be subsumed under the PDK (ex-CIPP) definition which requires that evaluation be "data gathering for future decision making." Now you might, if you have a copy of the great work in front of you, think that was a little unfair of me, because the basic definition of evaluation

which they use--"Evaluation is the process of delineating, obtaining, and providing useful information for judging decision alternatives,"--does not contain the word "future" at all. But we very quickly find from the way it is interpreted that "future" is the key point and the implicit definition in the PDK report is data gathering for future decision making. As a matter of fact, two pages before the definition there is a page which contains on it nothing but the following enlightening slogan, in capital letters, THE PURPOSE OF EVALUATION IS NOT TO PROVE BUT TO IMPROVE. Now that's great for formative evaluation, but that is not, of course, the same thing as evaluation in general. I don't think one wants to restrict the conception of educational evaluation to formative evaluation (they actually restrict it further than that). So it seems to me a mistake to try and tie it in to data input for future decision making. Evaluation suffered for a long time from being regarded as simply summative, but we don't have to swing so far over as to say it's <u>never</u> summative. We find many cases later on where it's even more obvious that they are thinking only of evaluation as data input for <u>future</u> decision making. I think the mistake here is like the mistake of defining government, for political science texts, say, in such a way as to cover only good government. What you should do, I believe, is to define government as neutrally as you can, and then get into the question of what

distinguishes good government from bad government under the heading of political philosophy; in this case the philosophy of evaluation-as-it-should-be-done by contrast with the definition of evaluation. This attempt to use a persuasive definition is something which greatly affects the whole treatment and I think is unfortunate. Another thing that it excludes for example, are evaluations done on the basis of instant gestalt-trained judgment. Now in many areas, for example in the grading of livestock and veterinary equipment, grading and evaluating in a straightforward way goes on and is essentially a perceptual activity. Probably the evaluation of students depends heavily on this instant gestalt person-perception activity. I don't want to exclude that by definition, but it isn't evaluative on their account. It's a judgment of worth or merit, and how it is done, whether that's valid or not, is a quite separate question.

And so my first worry then is that to use this particular account excludes some important types of evaluation, which you may wish to condemn as irrelevant or improper but I don't think you would want to regard them as not being evaluation at all.

The second problem with their conception is that it includes vast areas of the non-evaluative cognitive domain, e.g. administration theory and large parts of data gathering in the educational area, a decision which seems to me to unhealthily dilute the notion of

evaluation. The notion of evaluation essentially involves the judgment of worth or merit. Now to do an evaluation you've got to go and gather a lot of data first. It's reasonable enough, to say that that's part of the job of an evaluator, but it's confusing to suggest that doing it is a kind of evaluating. Since a theoretician also has to do just that, you might as well call it theorising. I think it's wrong to suggest that most of what they talk about as context evaluation is really evaluation. It is not; it's a market survey. After you've done the market survey, which you indeed must do before you can do a good evaluation, then you get started in the business of tying the needs you uncover in with performance criteria of other alternatives and making judgments of worth and merit and producing the evaluation, whether it's formative or summative. But I don't think it's helpful to talk about "context evaluation." It's true that one part of a market survey may sometimes involve an evaluation of competitive products, in the strict sense; but most of it--often all of it--is simply a survey of wants. And that part of a market survey is anyway not part of what PDK call context evaluation. They call that "input evaluation." But input evaluation as a whole might better be called a resources and options survey in which some evaluation of the relative merits and efficiency of those options and resources goes on. But a lot of it is simply survey.

Then process evaluation is--hard to work out, these are very, very tricky terms that are defined in six pages of closely written material; they're not defined briefly at all--process evaluation isn't quite, as I thought it was when I first looked at it, essentially formative evaluation. It is mostly monitoring and bookkeeping, two of the three elements they identify as process evaluation in this report--slightly changed from the standard CIPP account of this, I think. It may even include social bookkeeping. Now these may be tremendously important for a specific evaluation, they may be necessary for a school system's operation, but they are not themselves formative evaluation at all (contrary to Bloom's unhelpful bowdlerization). They may be <u>feeding into</u> one. But process evaluation, in the technical sense they use it, is by and large not formative evaluation at all (see page 315).

<u>Product</u> evaluation, surprisingly enough, turns out to be both summative and formative evaluation. I don't want to impose these terms of mine. I just mean by them the kinds of evaluation that are used (a) to improve a developing product, etc., and (b) to determine the merits of a completed, unchangeable one. The actual process of evaluation--the nature of evaluation in one sense--is usually the same in both cases, but the <u>role</u> it plays and, in a sense, the kind of entity evaluated is different. The feedback

loop from formative evaluation is within the project information flow chart--it terminates at a decision maker who controls the next R & D cycle, or someone who controls him. In summative, the feedback is to a consumer, typically, or a spectator (a historian, for example) or--perhaps--to a judge of the producer who may be considering hiring him for a related job. PDK's mistake is to take the decision-making role of many of these users of summative evaluation as definitional. But if the term "decision-maker" has any content at all, it does not include the fan in the stands evaluating a single play, or the academic version of him in the history department. (If you do call these "decision-makers," e.g. because they "decide" on what judgment to make, or because there are some actions of theirs that are affected by large numbers of these little evaluations, then you have totally diluted the PDK definition, since everyone is now always a decision-maker and the process of obtaining information for these decisions includes every kind of observation and reflection--in short, all cognitive processes are evaluation. The trivial or profound sense in which this is true eliminates the possibility of any theory of evaluation in the sense which PDK undertakes, since they do not include the whole of cognitive and creative psychology.)

In these terms, what is "product evaluation" as the term is used in the PDK report? They frequently say things like this: "This is what has traditionally often been thought of as all there is to

evaluation." So you think to yourself summative evaluation, in my terminology, but it turns out in the fine print that that's not true at all, since they insist that product evaluation itself must be part of the decision-making process: and the only kind of evalution of which that's true is formative. So, it seems to me, these categories are rather misleadingly referred to as four types of evaluation. I don't think that's the best way to put it, not that there's a sharp line between "gathering data as a basis for making an evaluation" and "evaluation," but just because the CIPP approach fails to demarcate the data claims and the value claims themselves.

To make a much tougher and more pedagogical claim about the CIPP analysis, it seems to me about the most complicated and confusing way of analyzing the practical procedures of evaluation that I can imagine, and it's certainly the most complicated one that I've ever seen. Not only is it impossible that teachers will grasp this, or that school personnel are going to use this without very intensive in-service training, but I think it's very doubtful whether what they'll do after substantial training will repay the cost of the training. I don't think we have to say that this is such a complicated subject as relativity theory where that situation would not be surprising. I suspect--to put it pragmatically--that ten to one condensation of CIPP would gain so much in teachability that any distortions introduced would be overshadowed by improved

comprehension. I think PDK have a duty to try this, or at least to bet that it can't be done, in which case I'll try it. I think it's terribly important to do this. The less jargon we can get by with the better; let's junk "formative" and "summative" and all these other terms, "instrumental" and "consequential" and so on, along with funny terms like context evaluation, and let's see if we can produce equally good evaluators in less time without them, or better evaluators in the same time.

I would like to go into details about the definitions of these terms, but time is short, so let me instead try to make some practical points. The first practical point is that <u>these</u> arguments about definitions are not "mere semantic issues" at all. Many programs are not getting adequate funds for evaluation, but those who run the programs and those who run the evaluations often have completely different ideas as to what they're supposed to do with those funds. And I don't think the PDK report is going to do enough to reduce this confusion because it includes too much and also excludes some aspects of responsible evaluation as I see it. There have already been cases where the granting agency terminated the evaluation contract because there was "not enough data-gathering" going on, although no case was made that the required further data was necessary for evaluation of the program in question. PDK is quite clear about this kind of point at times. At one place towards the end

they say, in effect, Don't fool around gathering detailed cost data if cost is not an object (in an experimental program, for example). Don't waste all that resource and time and effort and thought. All right, but if you take that seriously, then you've got to be much more precise about the distinction between general social bookkeeping or monitoring, and getting precisely and only those pieces of information that you've got to have for the evaluation. In particular, it should be clear to everybody remotely concerned with evaluation that one does not have to know a single damn thing about what is going on in an educational process in order to know that the project or method or process has completely failed, completely succeeded, or come somewhere in between. This is not always true, and that is why I say that the crucial point to understand is that the evaluation may require absolutely no knowledge of what went on between Day One and Last Day. Naturally, if there are process criteria in the criteria of achievement, you'll have to look at process. If, for example, you think it's important that the classroom be run democratically, you'll have to look in the classroom. But if you are using retention criteria you don't need process data (except to identify the occurrence of the experimental variable in the experimental group and its absence in the control group(s)). But USOE isn't too clear about this; at times they feel that if you're not collecting lots of data, you're not doing a decent evaluation. Even for formative evaluation,

this isn't true, where you often do want to know which features of an educational package were responsible for its success. To the extent you do this, you are going beyond evaluation in the straightforward sense. Evaluation of a package is simply doing what it takes to decide if the package did any good. Deciding what effects it had without using the language of merit is both a narrower and a broader enterprise than evaluation; and deciding what it was in the package of merit is, of course, an analytical enterprise of considerable complexity, properly called educational research but not--except by association--evaluation. Even deciding what the package was, that is, settling on an appropriate description for it, is process research. It won't do to argue that these things are all sharply separate--they're not. It won't even do to argue that they should all be made as separate as possible. For example, one of the most useful kinds of evaluation is the radical comparison; to take an imaginary example, one might assert that the talking page device does a good job of teaching certain vocabularies, but no better than the same page without the expensive talking feature at 1/9 the cost. That's useful evaluation. It's useful in the PDK sense: it feeds into a future decision. It is much more useful than just saying that the talking page does a great job. The talking page is a damned expensive item and the question that's important is whether that's where you ought to put your money. Now "radical comparisons" require some analytical comparative-fractionating research. So

I'm not arguing for a sharp distinction, but I am arguing for distinguishing whenever in fact the distinction doesn't cost you, because it will save you a great deal of money to make the distinction unless you actually have to run the evaluation into something else. For example, Sam Ball ran up the bill for evaluating Sesame Street quite a bit by getting into some research questions about which parameters controlled what variance. Interesting, useful, probably justifiable--but not in the guise of evaluation.

I'm pretty nervous about the rather casual way in which the PDK team dismiss what they refer to as the classical experimental design model for evaluation. They do not give detailed reasons for this. They say things like "what you need to evaluate a pupil may not be what you need to evaluate the utility of a program." Now that may well be true. But you need to get down to cases and say when it's true (if it is true) and what general conclusions can be drawn from that. I myself feel that the strength of the classical comparative type of evaluation we've always known is still very much greater than the suggestions by Cronbach and PDK would have us believe.

Turning now to a more serious point; this conception of evaluation has grave professional consequences for us, among them the elimination of fundamental criticisms of the client's objectives, since these are accepted as the axioms for the study if so presented

by the client. That is, unless he comes and asks you to help with these, to a large extent they are accepted. (Page references on this important criticism--and there are times when they jump the other way--include the following: 489 (a good clear one), 183 327, 410, 387, 411, 414, 419, and 422.) In my view, even when the client does not want his criteria criticized, the responsible evaluator is completely obligated--contrary to almost everything the PDK reports suggest except in a few places at the end--to subject them to the most minute scrutiny both before and after accepting the job. For these criteria may and frequently do contain unsuspected and well concealed anti-social, anti-personnel, impractical, or inconsistent assumptions, not to mention false and unclear ones. This fatal error of orientation shows up throughout the report, and it has deep philosophical foundations, as we see on page 41: "Selection of criteria always implies some value system, and values are essentially arbitrary even if not unreasoned." Now the cat's out of the bag. The fundamental approach of PDK is that value judgments are essentially arbitrary; and therefore you're not entitled to criticize those of the client. That seems to me a terrible position to adopt. It seems to condone an abrogation of the responsibility of the evaluator and I think it's precisely this philosophy that led to a lot of crap masquerading as evaluation that surrounded Title I. I don't think that it's arbitrary to conclude that Title I money was systematically

used throughout the South. It isn't arbitrary to reject a client's goals if they involve the misuse of funds appropriated by Congress for helping the needy. It is obligatory. The doctrine of equal rights in not, in my view, arbitrary but the most rational social strategy. Quite certainly, most evaluation today is simply naive if it supposes the client incapable of conscious or unconscious fraud on the government or the students or the parents or the teachers, and move one for an evaluator is looking into this possibility. (Move two is having someone else look into the possibility that he is himself putting one over.)

So it is not incidental but crucial that the evaluator is a conscience as well as a consultant, an auditor as well as an adviser. Whether in the formative or the summative role, he is worth little if he loses his independence. In summative evaluation it must be he who signs the report and edits it and in certain circumstances gets it published if the client refuses to publish it. (There is an echo of sensitivity to that point on page 301.) Otherwise he connives at preventable fraud. This is not a mere technical adviser's role, it is an autonomous professional role with a code of professional ethics attached, related in many ways to the auditor's role.

An amusing antecedent to this is to be found in medieval Japan, where great families depended for their livelihood and honor upon their leaders' skill and reputation as a sword-evaluator.

Their mark on a blade was a jealously guarded, hard-earned, and indispensable adjunct to the sword-maker's own signature. It was completely independent: whole families, generation after generation, depended absolutely upon the integrity of that signature. As a matter of interest, these early evaluators were strong supporters of the behavioral objectives approach. Criterion performance was checked out by the use of prisoners from the nearest jail, and the testers's signs on the blade indicated the "severage quotient" rather precisely. (Records do not show whether allowance was made for inter-subject variance.)

Well, this notion of independence is not clear enough in PDK, it seems to me. The very elaborate attempt, from page 470 on, to evaluate their own report, using their theory of evaluation, strikes one as a bit odd. Of course, we all do our own formative evaluation of everything we write, albeit not very well, but we can't possibly claim to do our own summative evaluation. The swordsmiths had the point right. They didn't sign the blade twice, the second time as an evaluator. They <u>broke</u> it if it wasn't a good blade. If they put it out over their signature, <u>that</u> meant that they thought it was a good blade. That's what one <u>expects</u> from reports, it's implicit in them. Not much can be gained from a second section which says "OK, now have we done well? Yes, we <u>have</u> done well." I think one just has to face the fact that there is a role of evaluation which it's just not sensible to suppose that you can do by yourself.

The practical issues connected with this point are numerous. For example, the institutional contamination of evaluators who are on the staff of school districts, projects or state departments, may make the PDK model for the use of evaluators in the educational system not quite viable. I do not think they examine with sufficient care the possibility that the role for such personnel is very tricky, and that much or some of their work is better handled by the use of outside consultants.

On a related point; it may be fatal to fragment the role of evaluation as CIPP and PDK do in their team approach. It seems to me they finish up without either an evaluator or an evaluation team. They break the task out into the statistician, the administrator, the public relations man, and so on. But you look at this list and wonder how this group is going to get together and produce an evaluation. It's not clear what they've got in mind, and I certainly don't think you want to add a moralist to the pack. But I do think they underestimate the role of axiological or value-analysis training for at least some of the team.

In conclusion, two minor points. A dimension of evaluation that I find of great importance in practice is the form of presentation of the evaluation itself. Quite commonly, compression--for example--is a merit that is hard to achieve and most important for dissemination. (This point is surely one of the reasons the letter-grade dies so hard.)

Now I want to make a correction of the handout version, some of you have, which was written for ERIC some time ago, from my notes without PDK's report to look at. It is simply incorrect as it stands. The PDK report does discuss mundane matters such as presentation, and indeed I am very glad they do. Something it does not do is to consider treating alternative forms of presentation as educational products themselves, and hence appropriate for evaluation in an experimental way, not by you as the author sitting there and saying you think they're good (having just written them), but by running a set of tests on them. So, for example, in coming up with the CIPP outline itself, why not treat that six pages or ten pages as being an educational product which should be evaluated? Perhaps in the workshop this was done, but it's hard for me to believe. To do it right, you get a naive graduate student from mathematics to read it, make what the hell he can out of it, summarize it in ten lines, and then try that summary as your control and see whether the teachers do better or worse from it. That would be a beginning on the "radical comparison" evaluation.

Finally, perhaps the best of all the practices of a good evaluator is getting the benefit of criticism from the other side. It goes very hard on the ego at times, but it's the mark of the professional to do it. I want to conclude by congratulating the PDK team for their willingness to invite the expression of deviant

viewpoints, not only in the use of an 11-man review panel in the formative stage but also in the use of this panel in the summative. I'm not sure they expected my viewpoint to be this ant, but arranging the panel was in the best tradition of a quite noble profession, and I salute them for it.

A last footnote: I think the list of specialties from which evaluators can learn, which they give on page 455, a useful kind of suggestion, could be supplemented somewhat. They say that evaluation specialists can receive "aid in comprehending their work from three reas: general systems theory, economics, and political science. The first is important to the evaluator as he considers the structure of phenomena on which decisions focus. The second, economics, sheds light on the nature of the decisions. The third, political science, contains constructs helpful in understanding the process of choosing." I would add these: (i) radical political theory--in order to see the other options in what they call input evaluation, (ii) ethics or value theory, never discussed by them--very important. They mention them once in one line very near the end, but they deserve better; (iii) Historiography--it never occurs to them that the analysis of historical material for evaluative purposes may have something important for us to learn from, and of course it undermines their definition of evaluation; (iv) accounting, a relatively unsophisticated approach to cost accounting procedures, cost-effectiveness

analysis, etc., is really important in some evaluations; (v) language analysis. In all their lists of specialties they never see that the skill of congruence identification between the description of goals and the items in the pool of tests is a skill in and of itself that is absolutely transcendent over statistical skills, test construction skills of the ordinary kind, and is the death of more test construction endeavors than anything else.

Well, one of the last things they say is "Nothing is worse than destructive criticism." I don't agree at all. Silence is much worse. I have at least avoided that.

DETERMINGING "MOST PROBABLE" CAUSES: A CALL FOR RE-EXAMINING EVALUATION METHODOLOGY

James L. Wardrop

It is illustrative of the scope of the PDK monograph that I feel after listening to these four expert reactions that have gone before, that I still have something to say. I have a feeling that if the members of the PDK Committee made a serious attempt to reconcile their presentation with our reactions, they would work for at least twenty years before coming out with a new one. To set the stage for my remarks (which Bill Michael has already alluded to), I should point out that one thing I have seen happening is people doing studies of a sort which would have to be considered as poor research and attempting to justify them under the heading of "evaluation." I would also like to say before I get into my substantive comments that although what I say might seem to imply that I am accepting the PDK definition of evaluation and the CIPP approach, this is not strictly true. I have used their definition and the CIPP approach because it provides me with a foil for my arguments. In this respect, then, I may not be quite fair to the members of the Committee.

My major point can be stated quite succinctly: namely, a central focus of educational evaluation is explanation (or, more

precisely, the selection among several possible alternative explanations).[1]

Once I have stated this thesis, the really hard work begins. I have now obligated myself to do three things: first, to explicate this succinct statement and attempt to give it substance; second, to justify the assertion I have made; and finally, to indicate in some way how my thesis may be viewed as a reaction to the P[illegible]terials. If I succeed in discharging any one of these obligations, my day will have been an unprecedented success.

A Notion of Causality

I circulated an earlier version of these comments to a number of friends, colleagues, and acquaintances (a few people actually fell into more than one of these three categories). A gratifying number of these people reacted. Now on the basis of those reactions I am convinced of two things: first, I had come up with the best projective technique for educational evaluators yet devised. I won't take the time to share with you those "projective" reactions right now, but let me say that if I wer to give you the list of names of those people who reacted and another list of the reactions, you would not have a very difficult time matching them. Additionally,

[1] I feel the need to qualify that slightly and say that although it is the central focus, it isn't the only focus and should not be the primary focus of many evaluation studies.

many of the reactions I received challenged my statement about the centrality of explanation to evaluation, for one primary reason. In that earlier draft I made the statement: "Explanation, as it is used in this paper, refers to the determination of the most probable cause for a phenomenon." To use the word "cause" especially with people trained in the social sciences, is either naive or foolhardy. It's sort of like sticking one's head into a beehive.

Nevertheless, I am going to stand by what I wrote then. (Naivete dies hard in me!) One difference, though, is that I am going to try to clarify what I mean by "cause" in the context of this paper. Ernest Nagel, in his chapter on "Types of Causal Explanation in Science" in Lerner's book Cause and Effect, has considered, among other things, what he referred to as "conditionally necessary causes." I'm not sure that I can do justice to that notion here, but I'll make a stab at it. That is, suppose event E was observed. When E occurred, antecedent conditions A, B, and C were present. (It is possible, as Nagel pointed out, that we may be unaware of the existence of some or all of these conditions.) The general rule which applies to this situation might be stated as follows: Given that conditions A,B, and C are present, then if condition D is also present, event E will occur; while if D is not present, E will not occur. Since condition D in and of itself is not sufficient to bring about the occurrence of E and since E may occur under some

other circumstances in the absence of D, we may speak of D as a contingently necessary cause of event E. This is precisely the notion of causality I had in mind in writing that "explanation--the determination of the most probable cause or causes for a phenomenon--is a central focus of educational evaluation."

It is my contention that, in every type of evaluation presented by the PDK Commission, explanation is crucial. Further, I would argue that the PDK volume does not adequately treat this concept nor does it adequately consider some of the implications of the concept for the methodology of evaluation. They have treated it somewhat at various points in the monograph, but nowhere is it presented in detail.

The Role of Explanation in Evaluation

In evaluation, as in experimentation, we seek to rule out, insofar as we are able, alternative explanations for phenomena. One aspect of context evaluation involves monitoring the system in order to identify problems and isolate possible causes of these problems. Since the subsequent delineation of a class of possible change strategies is directly determined by the causes so identified, it is vital that the evaluator be able to provide information of such quality as to insure that the identification of a cause or causes has a high probability of being correct. In other words, alternative explanations for the observed phenomenon or problem must be shown to be unlikely.

In input evaluation, also, the issue is one of explanation, the attribution of causality. For example, if we do something, A, then X will be more likely to occur than if we do B or C. I.e., A is a more probable cause--as cause was defined earlier--of X than are B and C. Once again, the decision (to do A, or B, or C) determines how and where and to what extent we are going to invest our resources. The ruling out of--or the assignment of low probabilities to--alternate explanations is critical.

One major focus of process evaluation is upon the early identification and removal of barriers to the success of the particular program selected to implement the change strategy. As before, we are faced with the need for valid explanations. To call something a "barrier to success" is to make a causal inference of the form: if Q, then probably not X. That is, the occurrence of (existence of) Q reduces the likelihood that X will occur (increases the likehood that "not X" will occur). Solving the problems of barriers is in this way formally equivalent to making the kinds of selection decisions which input evaluation serves, with the same implications relative to the attribution of causality.

Finally, product evaluation can be thought of as representing the effort toward final verification of the web of explanations which has preceded it. If the causal relationships postulated earlier have been correct (if the explanations have been valid), then the

hoped-for (intended) outcomes will occur.[2] It is in connection with product evaluation that we most often bring to bear the wealth of inferential statistical methods, apply our principles of experimental design, and in general call up our methodological "big guns." The concern in this paper is that we cannot afford to wait until this final stage to provide a sound methodological base for causal inference. The methodology of experimental design and traditional statistical techniques may not--and probably are not--appropriate throughout the evaluation process, but some methodologies must be employed which will provide us with a sound basis for our explanations.

The Search for Methodology

The preceding paragraphs have made a case for the centrality of "explanation" to evaluation as it is represented in the CIPP approach. On the basis of those arguments, one must conclude that the ruling out of (or assigning low probabilities to) alternative explanations--or at least providing data upon which to base such decisions about alternative explanations--is an important aspect of evaluation.

While the distinction between research and evaluation is important and needs to be emphasized (as the POK authors have done), I have

---

[2] It is appropriate at this point to remind ourselves that other, unintended outcomes will also occur.

some fear that a preoccupation with the differentiation may lead to an overly casual attitude on the part of some evaluators toward the quality of the information on which explanations produced within the evaluation setting are based. Threats to internal and--in some instances--external validity must receive extensive attention. If anything, they are even more important in an evaluation setting--where decisions (based on chains of causal inferences) determine the allocation of precious resources to a considerable degree--than they are in most research (especially basic research) settings. If a researcher commits a Type I error, he (or other researchers) may pursue an inappropriate question until the error is discovered and corrected. On the other hand, the possible consequences of an evaluator (or decision maker on the basis of information provided by the evaluator) committing the analogous kind of error are much more immediately felt in the resulting misallocation of resources.

The traditional model for educational research derives to a great extent from agricultural experimentation, after being filtered through experimental psychology. In his efforts to provide valid information on which to base explanations, the evaluator will often find this existing methodology both inadequate and inappropriate. In such circumstances, there are at least two alternatives to be considered. As a first step, we can seek methodologies for arriving at valid explanations which have been successfully utilized in other

disciplines such as sociology, economics, anthropology, history, and so forth. Tom Hastings, a couple of years ago in his presidential addre to NCME, addressed himself to this issue.[3] A second alternative, inadequacies in methodology have been identified, is to set out to develop new approaches for gathering and analyzing information, in order to minimize the probability that alternative explanations are in fact correct.

Identifying Methodological Needs

In the preceding section, I pointed out very generally a task for evaluation methodologists. One essential aspect of that task is the identification of evaluation activities for which existing methodology is inadequate. Through an emphasis on the underlying search for causality, we should be able readily to identify many of those inadequacies. This approach leads directly to a concern for the nature of evidence. What kinds of evidence will best enable the evaluator (or decision maker) to confidently discard alternative explanations as implausible? How can the evidence the evaluator collects best be communicated to the decision maker? There is also a question here and the adequacy of the explanation. One

[3] Hastings, J.T. "The Kith and Kin of Educational Measurers," Journal of Educational Measurement, 1969, 6, 127-130.

way of viewing that is to say that the explanation is adequate when the recipient of that explanation is satisfied. I think we need to distinguish explanation in evaluation from explanation in research at least partially on this basis. As an illustration, in a paper session on curriculum evaluation, one of the presenters discussed an evaluation of an approach to preventing dropouts in high school. In the course of the evaluation, he collected data on attendence rates for those students in the experimental program and students in some of the control groups. Having first noted that attendance was much better in the experimental setting, he went on to offer a couple of possible explanations. First, it was a work-study program and if a student did not go to school on a particular day he could not work that day (and therefore wouldn't be paid for that day's work). Secondly, any time a student was absent, somebody immediately called his parents--either at home or at work--to find out what was wrong. After the presentation somebody in the audience got up and said, "Wouldn't it be nice if we could get better information out of this kind of situation by designing a little experiment in which we have perhaps a two-by-two factorial design involving these two tactics. If you think of 400 students randomly assigned according to conditions, one group whose absence was handled with the phone call and not being permitted to work, another group only by the phone call, a third group only

by permitting them not to work, and the fourth group nothing, maybe we could learn something about which variable--not being permitted to work, or the phone call to parents--is the important one." The reaction was: "That might be nice, but I was an external evaluator for this project and it wasn't my place to redesign the project in order to serve this kind of research need. For the purpose of the evaluation it was sufficient to show that what was being done in this project did in fact have the effect of reducing absence." Without going further, one may ask which of these two factors is more likely to be the cause? At this point both of them were operating within the program, the evaluator felt that he did not have the right --or the need--to intrude into the operation of the program to the extent of suggesting this kind of more traditional research design. For the purposes of the evaluation, the explanation was adequate. For purposes of acquiring generalizable, scientific knowledge, it was not.

Given the position of the PDK Committee that evaluation serves the decision maker, other very important questions arise: for example, what kinds of evidence is the decision maker willing to accept as bases for his inferences? Another question is: Are these the kinds of evidence he should (according to some criteria) accept? The hope is that there is some commonality among decision makers in terms of the kinds of evidence they are willing to accept, that

the answer to this question does not depend entirely upon the idiosyncracies of the individual decision makers, that given certain decision settings and decision types, decision makers in common tend to seek certain kinds of evidence. Answering the "should" question will take much hard, logical thinking and--probably--years of investigation in an effort to validate the outcomes of that thinking.

Summary

Let me summarize now and say that if properly carried out, then, the task of the evaluator is in some ways much more difficult than that of the researcher. First, the evaluator finds typically himself working in naturalistic settings, settings in which many uncontrolled--and often uncontrollable--sources of variation are operating. He is placed in the position of seeking consistent covariation over time and over context, such covariation to be important datum for his attempts at inferential explanation. Because the consequences of decisions based on evaluation data have considerable implication for (and effect on) the allocation of resources, it is imperative that gaps in existing evaluation methodology be identified and some of those resources allocated to closing the gaps.

You probably have made an inference about my comments by now, one I would like to reinforce. (You have probably made several other inferences I would rather not reinforce, also.) Namely, I do not have any panaceas; I am not even sure where the answers will come from. But I would like to see more people spending time worrying the issue of evidence, explanation, and causality in educational evaluation.