DOCUMENT RESUME

ED 068 393 SO 003 190

TITLE	Evaluation and Development Blow Your Mind Conference. Final Report: Technical Paper Number One. Appendix C.
INSTITUTION	Colorado Univ., Boulder. Colorado Center for Training in Educational Evaluation and Development.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	Dec 70
CONTRACT	OEC-0-70-4770(520)
NOTE	38p.; Proceedings from the Evaluation and Development Blow Your Mind Conference, Boulder, Colorado, December 10, 1970
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Conference Reports; *Educational Development; Educational Improvement; *Educational Research; *Evaluation; Material Development; *Models; *Research and Development Centers

ABSTRACT

The Evaluation and Development Blow Your Mind Conference was one activity initiated under the design grant for the Colorado Center for Training in Educational Evaluation and Development (at the University of Colorado). Prior to the conference, 3 consultants (a psychology professor from the University of Washington, a representative of the Educational Testing Service, and the program director for the Bureau of Applied Research at Columbia University) spent 1 1/2 days evaluating the past and potential performance of the University of Colorado's Laboratory of Educational Research as a training facility in research and research-related areas. At the conference, these consultants were asked to free-associate and brainstorm on what evaluation, research, and development in education should be. This document is the lightly edited verbatim proceedings of the consultants' input and interactions with other participants. (GC)

Final Report

Technical Paper Number 1

APPENDIX C

EVALUATION AND DEVELOPMENT
BLOW YOUR MIND CONFERENCE

Colorado Center for Training in
Educational Evaluation and Development
University of Colorado
Boulder, Colorado
80302

The work reported herein was produced under contract with the U.S. Office of Education, Contract No. OEC-0-70-4770(520).

Final Report

Technical Paper Number 1

APPENDIX C

EVALUATION AND DEVELOPMENT

BLOW YOUR MIND CONFERENCE

Colorado Center for Training in
Educational Evaluation and Development
Universicy of Colorado
Boulder, Colorado
80302

The work reported herein was produced under contract with the U.S. Office of Education, Contract No. OEC-0-70-4770(520).

INTRODUCTION

On Thursday afternoon, December 10, 1970, the Evaluation and Development Blow Your Mind Conference was held at the Red Lion Inn, just east of Boulder, Colorado. The Conference served as one activity among several initiated under the design grant for the Colorado Center for Training in Educational Evaluation and Development. The Conference was coordinated by Dr. William L. Goodwin, the Design Project Director, and was attended by a hearty (considering the inclement weather) group of 26 persons representing a variety of organizations, as can be noted below:

1) Biological Sciences Curriculum Study, Boulder, Colorado:
Dr. James T. Robinson, Consultant.

2) Colorado Department of Education, Denver, Colorado:
Dr. Arthur R. Olson, Director, Assessment Evaluation Unit.

3) Denver Public Schools, Denver, Colorado:
Mr. Barry Beal, Supervisor.
Dr. Jerry Elledge, Supervisor.

4) Earth Seience Educational Program, Boulder, Colorado:
Mr. Larry Irwin, Associate.
Dr. William D. Romey, Director.

5) Interstate Education Resources Service Center, Salt Lake City, Utah:
Dr. Brent Gubler, Director.

6) John F. Kennedy Child Development Center, University of Colorado Medical School, Denver, Colorado:
Lila R. Wegener, Intern-Trainee.

7) Social Sciences Education Consortium, Boulder, Colorado:
Dr. Irving Morrissett, Director.

8) Southwest Cooperative Educational Laboratory, Albuquerque, New Mexico:
Dr. James C. Moore.

9) Southwest Regional Educational Laboratory, Inglewood, California:
Dr. Mos Okada.

10) University of Colorado, Boulder, Colorado:

Fellows and Students of the Laboratory of Educational Research

Beverly Anderson
Richard Bennet
Evelyn Brzezinski
Nancy Burton
Arlen Gullickson
Norris Harms
Larry Nelson
Susan Oldefendt
Rory Remer
Todd Rogers

Faculty

Dr. Thomas Barlow, Associate Dean of the Denver Center School of Education.

Dr. Gene V Glass, Associate Professor and Design Project Staff Member.

Dr. William L. Goodwin, Associate Professor and Design Project Director.

Dr. Kenneth D. Hopkins, Professor and Design Project Staff Member.

Dr. Gerald W. Lundquist, Assistant Professor.

Playing principal roles at the Conference were three consultants:

1) Dr. Arthur Lumsdaine, Professor of Psychology, University of Washington.
2) Dr. Sam Messick, Educational Testing Service.
3) Dr. Sam Sieber, Program Director, Bureau of Applied Research, Columbia University.

These men spent Wednesday, December 9, and Thursday morning, December 10, evaluating the past and potential performance of the Laboratory of Educational Research as a training facility in research and research-related areas. On Thursday afternoon, the consultants were asked to free associate and brainstorm on what evaluation and development in education should be like and, consequently, on what training experiences evaluators and developers should undergo.

What transpired is recorded, for the most part verbatim, herein. The consultants were sent transcripts of the Conference and asked to edit sparingly (in order that the spontaneity and original flavor of the Conference thereby might be preserved). It is obvious that if the consultants were writing on (rather than discussing) the same subjects, their products would be somewhat more polished and organized. Still, the contributions made by each of them seemed perceptive and noteworthy, and are presented here as inputs to be considered in this general area.

THE EVALUATION AND DEVELOPMENT
BLOW YOUR MIND CONFERENCE

LUMSDAINE: This is really a very unstructured assignment, although it has some precedents. I think that among us there have been some past expressions of disagreement with respect to the relationship between research and evaluation and their relative importance. I guess I should identify myself as a heretic and, furthermore, a renegade because I was raised as a basic researcher. Starting with conditioning under Jack Hilgard at Stanford, I gradually progressed through a series of educational research studies in the classical hypothesis-testing paradigm. I have arrived at a point where I seriously question whether the most fruitful way to proceed, in terms of improving education, is in fact through the classical (if I can use that phrase) model of nineteenth century physics: one goes into the laboratory or equivalent and does basic research; from that derives implications which are supposed to say something about what educational practice should be; next one does dissemination or diffusion or something like that and these basic principles somehow translate themselves into practice.

I'm sure that this happens to some extent, but I guess that the position that I would take (at least for the sake of argument and really a little bit more than just for the sake of argument) is that the educational research and development dollar could well have a larger portion of itself expended on something more like an engineering model than a science model. That is to say, a development, evaluation, and test model, oriented around the development of educational products or programs, the empirical testing of them, the use of data to improve them, and the arrival at general findings arising in the context of this quite frankly applied, engineering-type development.

I hope that the term "engineering" doesn't turn you off; what I really mean is the attempt to develop and improve (through research and applied-research techniques) a useful educational product with concern always, of course, as to the extent to which what is found in one situation can be applied next door, next year, across the country in a slightly different context, and so on. Well, we have had several spirited interchanges, particularly Sam Messick and I and a few of our other colleagues, on this general proposition; and I think that maybe the stage could be set (entertainingly if not usefully) for some further discussion by asking Sam to identify some of the points of disagreement that he perceives.

MESSICK: Okay. It's a little bit difficult to know how to proceed at this point, but let me make a few general remarks, first by starting out specifically, then becoming more general, and then, hopefully, coming back to the more specific again. Specifically, I would like to take exception to the engineering paradigm as a way to proceed in educational evaluation and development, being deliberately a little bit unfair to Art in this position and setting up somewhat of a straw man. However, I'm also setting up a straw man because I'm concerned that the adoption of an engineering paradigm in educational development and evaluation (although technically it is feasible for us to worry about generalizability of the findings next door and next year) would make such generalization unlikely. Rather, I feel that we should be adopting a paradigm that puts much more emphasis on process, that is, concern not with just assessing the size of effects and finding them good or bad, but concern with understanding the processes which produced those effects. The engineering paradigm in its simpliest form is essentially concerned with input-output differences relative to cost; and I would argue that that's not enough. We have to be concerned with the context in which the output differences occur, with the processes that produce the outputs, and, notably, with the antecedents and consequences of the total enterprise.

Let me become a little bit more general at this point. If we ask what educational development should be and what educational evaluation should be, it seems to me that we can't proceed very far in answering those questions without asking, "For what?" Educational development for what? Educational evaluation of what? Very early in the game we have to worry about the goals of the educational programs. That is, educational development and educational evaluation cannot be considered separately from educational programs and their specific content. Educational programs cannot be considered separately from the goals or values of education. I argue strongly that none of these things can be considered separately from educational research. I argue that evaluation of educational programs ought to be research on educational process and nothing less. And to a very large degree, educational development should be the same thing.

At Educational Testing Service, we have several divisions with somewhat differently articulated missions. We have a division that is called the Division of Test Development; I'm sure you all know from it's name what it's supposed to do. What is test development? What we consider test development to be might give us some lessons as to what educational development might be or mean. At Educational Testing Service we also have a Division of Developmental Research; it's not so clear what this division is supposed to do. In our Developmental Research Division, the concern is to try to understand the dimensions of a problem and the variables that are influencing an educational problem in order to develop options for solving that problem and then evaluating those options with respect to their relative effectiveness. Thus, the task is trying to understand the nature of the variables and doing the hard intellectual work that's involved in construct validity research in order to develop test specifications of the processes to be assessed, the specific content tc be assessed, and whatever other dimensions are involved. Once those specifications are clearly in mind, and maybe a prototype item or two has been developed, then the Test Development Division comes into play. The Test Development Division, then, is a group of subject matter experts that develops items to meet fixed specifications. If you don't know what the specifications ought to be, then that's not development in our particular ethnocentric view of the problem.

As we consider the process of educational development, I think we have to worry about what we mean by that process. Is it the generation of products to meet specifications that are well understood or is it an attempt to understand what those specifications ought to be? I would argue that at this stage of development, we have to understand what the specifications ought to be. At this stage our concern is with educational evaluation; we have to understand how effective the approaches are and understand the processes that produce those effects. I view both of those enterprises, both development (as I've talked about it) and developmental research and evaluation, as evaluative research. From my point of view, we should maximize the overlap between evaluation and research and between development and research.

This is not to imply that other approaches to development might not be required because of the press of circumstances; that is, we might have a problem facing us that must be solved immediately. For example, the social fabric of the schools is disintegrating; we can't take five years out to engage in a research program to understand how to best proceed in the future. This problem and others have to be addressed now. One consequence of this type of pressure is that we might develop programs which are not researched, programs where we don't pay any attention to understanding the basis for their operation.

By the way, there's another subtle political pressure that is operating presently in Washington. This pressure is a call for research on the excuse that we don't know enough to act. Just two years ago the pressure was exactly the opposite; then it was a call for action, labeling research as a frill. But now it is a call for research because the knowledge base isn't substantial enough to indicate how to act. The reason for that switch is very clear to me; I may not be right, but I think I perceive it clearly. The reason now that there is a call for research is that research is really very inexpensive. You can do a lot of research for $300,000; you can't do much action. So the call for research is essentially an excuse for not supporting important action programs.

But I don't see this as a dichotomy; I don't see why we have to talk about a pendulum swing. There is a viable alternative, that which I called evaluative research (and that Kurt Lewin 30 years ago talked about as action research). This strategy is used when, faced with pressing social problems which must be addressed now even given all of our ignorance on these issues, we develop action programs based upon the best available ideas and knowledge, and we do it now. We attempt to implement those programs immediately—we don't wait for five years of research. Additionally, we build into the implementation of those programs provision for collecting information, relevant information as to how the program is operating, information about whether or not effects are being produced, the size of the effects and information about the processes that produce the effects. That makes the action program an action research program. I would argue that action programs without that research component are a waste of time because it is unlikely that we will understand enough about unresearched action programs to generalize therefrom to other settings, to other individuals, or to other times. That's a fairly drastic way to present the case but nonetheless I think that a kind of (if you will pardon the expression) "black-white thinking" would warrant it.

SIEBER: Let me just state my position on this very briefly; I think Art wants to rebut. I agree with Sam that evaluation should really be looking at the process, the procedures by which you get the outcomes. Otherwise, you cannot generalize the results of the evaluation to other settings, to other types of students, to other what have you. In that respect, I think I disagree with Art on the emphasis on the engineering-model product evaluation. But then I want to carry that implication further: What does that do to our research design? If we have to look at every possible condition, constraint, facilitating situation, and procedure involved in every kind of program or educational practice (and being a sociologist I'm more concerned with the evaluation of larger scale programs, organizational changes, etc.,rather than particular little changes in curriculum, in a certain grade, for example), what does this do to our emphasis on quantitative experimental design? If your evaluation considers processes, you have to look at everything. You have to take into account the SES background of students (maybe even of the teacher), the teacher's personality, teacher enthusiasm for the new program that you are inserting, how long they have been in that school system, the school system's innovative climate, the principal's backing up of the program, the amount of pressure on the school to perform well in this new program, etc. All of these things might be contributing to that outcome you are getting. You are not going to be able to design an experiment which is sufficiently elaborate unless you have most of the schools in this country, I think, to control for all of these procedures, constraints, opportunities, and so forth, in that situation. So I am wondering if this will necessitate a shift to more qualitative types of observation, say impressions, rather than attempting to quantify the outcome of every variable in that situation that you're interested in studying.

MESSICK: That's a very good point. Essentially you are saying that in education, as we consider the effectiveness of a particular program or product, we have to recognize that it's occurring in a context which is essentially a system. It is a very complicated system because it is a system that is addressed essentially to the whole problem of human development. We recognize that educational growth and the understanding of educational growth can't really be separated from the understanding of human growth. The system we are dealing with in education embraces not only the child, his teacher, and the program, but also the home, the school, the peer group, and the community.

In order to deal with that problem properly, we have to address the possibility of interactions occurring and that means we must utilize multiple measurements. There is no alternative to that as I see it. The alternative of simplifying reality by controlling variation is one possibility (of going into the laboratory and making simplified but hence artificial situations that hopefully will help us understand part-processes), but not a particularly good one. That is, we may understand part-processes generated in the lab which, when we try to generalize to the real world, are found not to generalize very well. Thus one consequence of the point of view being advanced is that educational evaluation and educational research must be multivariate and must be interactive. I would also argue that it must be longitudinal and comparative and that if the research enterprise is not longitudinal (in the sense of having multiple measures over time), we can't understand the processes. Research which is multivariate, interactive, and longitudinal is very complicated and must involve very complicated designs and very complicated multivariate analysis.

SIEBER: I think we're speaking of designs that we really haven't thought through yet ...

MESSICK: Right.

SIEBER: Just haven't arrived at yet. "Strategies", rather than "designs", may be a better term.

LUMSDAINE: I want to go back to the remark before last before last.

MESSICK: The ante-penultimate remark.

LUMSDAINE: I thought that I was going to pick a fight with Messick and I found myself discouragingly in agreement, from one point of view, with what he said. I guess that maybe I never should have used the term "engineering" and maybe some of those polysyllabic adjectives; this kind of term, if we aren't careful, may do more to obfuscate than to clarify. Since there isn't time for jokes, I thought that I might read a little thing that a friend of mine handed me the other day; some of you may have seen it. It says in large type, "No wonder we don't communicate" and under this in smaller type it says, "You say you understand what you think I said. What you don't realize is that what you heard is not what I meant."

When I talked about an engineering model, I was really saying, or trying to say, very much what you subsequently said was what we ought to do. That is to say, you're saying that the press of events, the need for educational improvement and innovation is such that we can't possibly do all of the background research necessary. Rather, we have to, in fact, take action; that is Point 1. And I was really trying to say that same thing. I despair of trying to do all of the fundamental research and then deriving from that the principles necessary to make the decisions on what kind of action programs we are going to engage in to meet pressing educational, societal problems. So, we're in agreement on that; that some action, some development, is needed with at least comparable, and I think maybe considerably higher, priority than the more traditional on and on and on with fundamental research (out of which we trust, perhaps rightly, that ultimately some good will come).

But then the next thing that you said, seemed to me, was that just development, just innovation, just new educational products or new gimmicks (if you want to use a pejorative term) or just new methods or approaches aren't going to help much, aren't going to lead us anywhere very much, unless we combine that with "action research" (I prefer to call it "evaluative research"). That is to say, evaluative research which is frankly applied in its orientation, which seeks to determine what a particular product or program accomplished, and how that compared with some reasonable alternative which has enough stability that we can define it, so that we know what we are comparing.

This is exactly what I meant by an engineering approach. I meant to say, first let us see what the problems are (whether they're those of elementary reading or of getting a more intellegient non-bomb-throwing type of participation by college students in the political process or whatever they might be). Let us try to devise and implement programs that seem to meet the needs that face us, and then let us address a very substantial amount of our research effort to determining whether in fact the kinds of outcomes that we hoped these products would bring about are, in fact, brought about. However, we then need to go one step further because history never repeats itself exactly. What we find out about a particular product designed or a particular program implemented in Peoria in 1969 doesn't necessarily give us a completely secure basis for claiming that we can transplant the same thing to St. Paul or Austin in 1971. Well, it seemed to me, though, that you were saying that the way we get this assurance of reproducibility (that isn't the term you used but it's the one that I'll use), of exportability, is in terms of the evaluation of that particular approach. (I mean certainly to include hard-data evaluation to assess the extent to which outcomes demonstrably are realized not just whether the new product or program sort of looks nice.) How do we get this assurance that what we find has been accomplished by a particular program, reading program, mathematics program, or what have you, in a particular place can be depended on to produce similar effects in some future situation? What I further heard you saying (and maybe I was reading too much between the lines) was that one of the ways that we do this is to look at the process, the fundamental variables that seem to be involved, the many respects in which a new situation can differ from the original situation, and try to see if we have a theoretical basis for determining whether the proposed generalization to a new situation probably is secure or whether it is risky.

I think, from my particular biases gleaned from the kinds of things I was doing for a fair number of years, I would say there is another basis for reproducibility, and that is the extent to which the method is embodied in a concrete set of materials and procedures which are physically reproducible and exportable. If you take the so called "programmed" materials, or film programs, or video-taped programs, or computer programs, they suggest programs that are embodied to a considerable extent in things which are inherently physically reproducible in terms of educational media, then that is another basis (not the only basis but another basis) for trying to assure reproducibility. At least you have greater assurance that the educational stimulus will be the same in the new context once the program is transplanted to a different state or a different school system, than you would have if the basis for similiarity and reproducibility is only some kind of "method" stated in abstract terms that is diffused by a type of multi-stage diffusion process such as teaching it to teachers of teachers of teachers.

So I think this is another basis for exportability which needs to be given serious attention. In terms of the development and evaluation paradigm, the development of specific products of this kind has a concreteness (that again suggests what I thought I was trying to imply by the engineering model) that just the diffusion of methods or the dissemination of principles of methods does not have.

MESSICK: I think the danger here is in taking you too literally on the emphasis of products; that is, there are really two problems that we're concerned with. One is the problem of generalizability or reproducibility that Art has emphasized. The other one is the problem of interpretability. It seems to me that we must be concerned with understanding the basis from which the effect is produced and attributing that effect to an appropriate kind of treatment variable. In the current engineering model, emphasis is on how you get to the output from the input (that is, the outcome-input differences) and (as stimulated by Washington) on cost-effectiveness; we have to consider output relative to cost. Rather than the traditional engineering model, I would prefer to see used a model which is really a kind of a way of thinking, a way of emphasizing certain problems, a model which Micheal Scriven alluded to ... just kind of tossing it off in a sentence, in an article ... saying that really a better model for educational research (for me that means educational evaluation and educational development) is a medical model. That is, a model concerned not with just input-output differences (but one that is concerned with those too, to be sure), but also with such things as side-effects. The engineering model, to me, faces the danger of emphasizing completely intended outcomes and then evaluating effectiveness in terms of the extent to which intended outcomes are met without being much concerned with unintended outcomes or side-effects. The medical model, it seems to me, is a much better model to guide our thinking, say, in the evaluation of drugs. It is just not enough to evaluate a drug by saying it produces the intended effect. Here's a drug that was supposed to reduce blood pressure and it does. If it also disintegrates the liver in the process, it's not acceptable. So we must be concerned with side-effects in that regard.

I think it should be clear that there is controversy over this point: that there are people who will take the other position and say "no". We, Art and I, were at a conference together a couple of years ago in which one of the participants essentially stood up and said "no". He indicated that he was a member of the butter-wrapping school of evaluation.

By that he meant that if you want to teach a person how to wrap butter, then only one thing is important in evaluating the training program, and that is to assess how much butter trainees wrap, and that's all. It's a very, very simple way to evaluate the effectiveness of the training program and the educational procedure used. Now I think it is easy to counter that approach. One way to counter it is to say, suppose there is a butter-wrapper who is the most efficient in the group, but who has gained his efficiency because of a particular process he engages in, a stylistic quirk, as it were. That is, he's very rapid because he touches his thumb to his tongue as he picks up the wax paper. Now you may not think that that is an acceptable mode of butter-wrapping, given hygienic standards, but he produces more wrapped butter than anyone else. There's a very important lesson to learn from that ... what it implies is that you may not be able to evaluate the desirability of the outcome without knowing and understanding the process which produced it. To understand the process which produced the outcome is an enterprise in educational research which I see as paramount in the medical model approach I suggested here, and considerations of evaluation and development are part of that model's research process.

LUMSDAINE: I knew this was going to be a rap session, but I didn't know it was going to be a butter-wrapping session.

GUBLER: Now what paradigm do you use to evaluate research? I think that's really what I hear all of you struggling for.

MESSICK: This raises another issue.

GUBLER: But isn't that basic?

MESSICK: Yes, it's an issue that we're going to get to. I think it might be well to raise it now. If you want to say what educational evaluation ought to be, you first have to determine what kind of things you are going to evaluate; then I'll tell you, maybe, what it ought to be. And if you ask what is educational development, what should it be; well, again, the question becomes what kinds of things do you want to develop? How do you answer that question? What do you want to develop in education? Very quickly we get to the whole problem of the goals of education. What is it you want to teach young children in the first grade? Or consider a problem that many people are discussing currently, what should be the content of preschool education programs? What should be the goals of preschool education? What should we teach the very young child prior to his entering the formal school system? And how did we ever decide, in the first place, what to teach young children?

It's very clear to me that the answer to that question is two-fold ... that there are really two issues just as there are two when we are asked, as scientists, to make a recommendation for practice. In the area of measurement, for example, if someone asks you whether they should use a particular test for a certain purpose, there are always two issues involved. First, is the test any good as a measure for what it's supposed to measure? The second issue is: should it be used for the intended purpose? Well, the same is true in the educational realm, more broadly. Is the procedure any good for bringing about the effects that it's supposed to bring about and should it be used to bring about those effects? The first question is a scientific one. In the measurement area, we have standards, psychometric criteria, for evaluating the adequacy of a measure, the most important of which is construct validity. The second question is an ethical one. It can only be answered in terms of the social consequences of applying the technique, the measure, or the educational program, and an evaluation of associated social consequences in terms of value systems.

The problem that we're in as scientists is that we frequently delude ourselves into thinking that answers to the second question can be obtained from answers to the first question. It doesn't matter how good a test is or how effective an educational program is per se, for answers to the second question; the second question can only be answered by appeals to values and ethical evaluation. When we ask how do we evaluate educational research, development, or evaluation, it always has to be for what? We are at a loss as scientists to answer the "for what" question because we really haven't developed many of the techniques that would help.

However, I do agree with Michael Scriven that science has a lot to say about answering ethical issues. Scriven says (and I think I agree with him, although I'm not sure) that ethics is a social science ... that the methods of empirical social science can be brought to bear on ethical issues. This is particularly true with respect to the very critical aspect of ethics which is to evaluate the consequences of alternative approaches. I think, clearly, that this is a possibility, but in the long run, the ultimate decisions are going to be made in terms of judgment and in terms of value. We have difficulty dealing with that in a pluralistic society because there are different values and different opinions about what is good for young children.

Right now we kind of cop-out on the problem of preschool education, which was the specific illustration I started with, by saying that whatever we give to very young children before they enter the formal school system should be good for them in the first grade. That is, whatever it is we are doing in the first grade, we should do it earlier. But that really is a cop-out because we should ask, "why is it good for them in first grade?" Is the answer, "because they need that kind of preparation for second grade?" And why do they need it in the second grade and so forth; you can see what's coming. Ultimately we end by saying that what they need in formal schooling is adequate preparation for effective adult-role functioning. That means that we have to turn to the nature of society and the nature of changes in society and the kind of adult-role function for which we would like to prepare children; and we are confronted with the fact that maybe we're not doing such a good job of that.

I heard a story just the other day which produced a great deal of negative affect in me. It was a story of an Indian tribe in Central America which, because of a lot of ecological constraints, had stayed in the same place for years and years and years, literally for centuries. Because of mountains that enveloped them and other factors, they just remained there. But over the centuries, they were subject to a parasitic infection which consequently caused blindness as individuals aged. So most of the adults in the community, age 35 or 40 or older, were blind. There grew up in this setting a non-formal educational system. It was essentially geared to preparing the youth of this community for the inevitable blindness that they would have to face. Many of the young in this country feel that that is exactly what the educational system is providing for them now; they're telling us that in no uncertain terms. Now how do we respond to that? How can educational research or development or evaluation respond to that in meaningful ways? I'm not going to answer that question for a very good reason.

LUMSDAINE: I think that we have some sense that maybe some of you in the audience would like to ask some questions; I have 10 other comments that I feel it's essential to make, but I am going to suppress all of them.

SIEBER: I just thought that it might be a way of interconnecting several points that have been made to say that evaluation is not context-free. Over and over, we keep referring to the context of evaluation in several respects. We were talking about the goals of, and the specifications of, evaluation. Sometime ago, at the beginning of this presentation, Sam stated that at ETS once they have worked through the specifications, they can turn the hack work of development over to the developers in that division. Specification is the big problem, and I think that it should be a big problem for any sophisticated evaluator who goes into the school. It can be a big problem in that the practitioner does not know what the goals are. You just can't take it for granted that he knows exactly what he wants to get out of a new educational practice or program. Even if he does know, you might spend a lot of time trying to help him articulate and specify those goals and, of course, it is up to you to operationalize them in some way. It also might be the case that even if he states specifically what his goals are, those really aren't his goals; he's misleading you or he's misleading himself ... the goal of adopting the program might be to improve community support for the school, to resolve conflict among school board members over a particular segment of the educational program, to get him a promotion ... you don't know what. So the evaluator often should diagnose and go behind the problem that is presented to him, even if it appears, on the surface, to be a clear-cut problem. Sometimes the practitioner gives you too many goals, and it's impossible to fulfill all of these goals in any kind of program; so you have to make him pare them down. The entire process of specifying and articulating goals, defining them, and diagnosing them is a very important stage in evaluation which has only been alluded to, I think, by Sam who talked about objectives.

He also talked about goals, I think, in the sense of policy research: should we uniformly accept the goals? Even if they are good goals, rational goals, and would help children learn a particular kind of curriculum, maybe there are more important goals that practitioners should have; if we supply them, we are inserting our own kind of judgment in the situation. Perhaps we should as generalists in education (as sort of mini-educational statesmen, that is, small educational statesmen). Maybe we should be thinking about what inner-city schools really should be doing for children. Rather than trying to develop a program that is supposed to teach them to read faster, maybe we should be saying, "wipe out the lecture" or introducing an entirely different kind of educational system or having kids write poetry or learn history by painting or learn how to read by memorizing scripts for plays ... something like that ... which may never have occurred to the practitioner. So what I am saying is that this is one situation in which context is important. What is behind the goals, behind articulating the goals, etc.?

Another place that context comes into play, which we also talked about in the very beginning, concerns all the multifarious variables, the restrictive conditions in the situation, the facilitating factors, that are producing the outcome ... those factors that we should know about if we are going to do process analysis and generalize and reproduce the outcome or modify it for other kinds of settings.

Another way in which the context comes in, is in the utilization of evaluation. I think we are all taking for granted (and necessarily just to keep ourselves alive and keep our egos strong) that it makes a difference whether we evaluate or not. But I think very often, and maybe in a majority of cases, evaluatión doesn't make any difference. In a certain sense, often our evaluation results, no matter how beautifully developed and presented, just have no impact on any situation whatsoever. The report is filed away someplace in the USOE or in ERIC, or it never gets to practitioners ... they couldn't care less even if it does get to them, they have to see it in action ... you know, a whole multitude of reasons — bureaucratic restraints, "it's going to cost too much, etc." — so that very often evaluation leads to nothing, absolutely nothing. One reason it doesn't lead to anything is because the evaluator never sits down with the practitioner in the beginning to talk about how the evaluation is going to be used later on, to get any guarantees of further testing-out of the results of the evaluation, or to determine how feedback will be provided to people who will be affected.

These are some ways, I think, that we have to broaden our conception about evaluation, and they certainly have implications for the training of evaluators.

MESSICK: Sometimes the evaluations are deliberately misused for political reasons.

LUMSDAINE: Sure.

MESSICK: We as evaluators will become pawns in the political process.

LUMSDAINF: Let me interject one dissenting note to this. One way in which you can make quite sure that the results of evaluation are in fact utilized is when we are talking about what Scriven refers to as formative evaluation, that is, if we think of a developer-evaluator team in which the customer for the evaluator is the developer himself. This is not a matter of passing judgment on something, of making an external administrative decision. Rather, it is the use of evaluative data, say in the case of a program on history, to find that certain concepts got across pretty well and that certain others didn't. Then you use this information (this is a simplistic example, but will do) to decide what parts of the program you should pay attention to in trying to revise and improve it. In fact, you virtually guarantee in a limited, perhaps, but still I think quite real sense, that the results of evaluation will be utilized. Further (again in what is a limited but I think important sense), it is possible to ascertain whether or not in fact you did what you were trying to do. There are a number of instances that I could cite on such use of evaluative data (that is, the use, in revising a program, of hard test data on what aspects of the intended, and if possible even the unintended, outcomes of an educational, instructional, informational, or indoctrinational program got across with an early version of it, and which ones did not get across).

In these cases, the data have been fed back in one or more stages of revision in which the revised product has then been tested comparatively against the original unrevised product, and where it has been extremely clear that the use of evaluative data in revising the program has led to a better product than one started out with. This has been shown even where the original product was one that already had been through considerable development and evaluation of a qualitative, non-data-based form. So this is at least one important exception to the pessimism about the usefulness and use of evaluative data.

MESSICK: I believe it is also important to consider the role that educational researchers and evaluators, and social scientists in general, are being called on to play in the political enterprise. That is, many very large scale evaluation studies have been undertaken in the past few years, primarily to gather information that would convince politicians that certain things are worth investing money in. Other situations are such that it is almost impossible to gather relevant information about the effectiveness of particular programs because of lack of foresight. Title I, for example, put millions and millions of dollars into the educational community. Since there wasn't any adequate pre-test information available, it is very difficult to evaluate, on a national scale, the effectiveness of those dollars. Other programs that are smaller, like the Headstart Program, have been evaluated in a variety of ways. Very recently a large scale national impact study of Headstart was undertaken by the Westinghouse Corporation; the constraints of the study were such that people who knew the manner in which the study had to be conducted, i.e., social scientists (by and large uniformally I think) predicted that the results had to be negative. That is, the results had to come out to make Headstart look harmful, or at least inadequate. Yet the study was undertaken, those results were obtained, and they had political consequences.

How could we as social scientists somehow have prevented that from occurring? Standing up and saying, on professional and scientific grounds, we feel that that evaluation should not occur in those terms because it's unfair, we know in advance the political consequences that are going to arise, and we do not view this as a scientific enterprise. And yet, again, this study was undertaken, but we were not organized enough or in agreement enough to take a stand that would have had political power in that regard. So, in a very real sense, we are pawns of the political process just because we have no control over the use of our evaluation reports or in the interpretation of them.

GUBLER: Should you have?

MESSICK: Well, that's another question. It depends on whether you like the results or not. If they're to be misused, then you stand up and say, "You are misinterpreting the data." They don't believe you or they don't take your suggestions into account. Sometimes, the political decisions are made before scientists are even aware of the data; the Westinghouse report was leaked politically prior to any opportunity for scientific review and evaluation. It affected the political climate negatively before there was any feedback available.

GUBLER: This is the dilemma we're in; that's why I asked my initial question. Ultimately you are asking what criterion you use to validate research. You can use the qualitative approach, I presume; you gave an excellent example of people establishing an informal educational system to teach and prepare for blindness which really wasn't needed. I think historically you can denote statisticians who have cited figure after figure, for example, justifying the need for more dollars to support welfare when we really ought to be trying to determine what is causing welfare. You know, this kind of a thing. So now we are back to the affect or the value domain, and we can evaluate that in terms of a group of specialists, scientists, or professionals or we can talk about a democratic process which operates on the majority principle.

MESSICK: We'd like to have a system where the democratic process may be applied in an informed way. Let's go back to the Westinghouse study, for example. When that was undertaken, it was undertaken as an overall global summative evaluation of the effectiveness of Headstart at a time when that was not the political issue at all. That is, the evaluation wasn't pointed toward the issue at hand. The political decisions about Headstart essentially had been made. It wasn't any question about whether it was any good or not; that was like asking whether parks are any good. How do you evaluate whether parks are any good? Headstart, for better or worse, for good or ill, was a remarkable political experiment in the sense that it converted a clientele into a citizenry. Because of the political power of that citizenry, Headstart was not going to be terminated. The important question was not whether Headstart was any good but, rather, what are the aspects of the preschool program that produced differential results. What can we do in preschool programs that will foster growth on cognitive, personal-social, and affective dimensions? The Westinghouse study didn't really bear on that issue. It kind of muddied the waters, in a way.

Let me try to present the problem in a somewhat different way. I am really concerned with the role of social science in the political process and want to point out and emphasize the fact that educational evaluation is in the political process whether we like it or not. Being pessimistic considering the present political climate, it is more likely that evaluation results will be misused than used properly on the national scene or on the state level. A friend of mine, a sociologist, posed the problem in the following way. He said that whenever there is a complicated enterprise in which many parties are involved and the enterprise fails, the finger of blame is always pointed toward the weakest party. For many many years that finger of blame was pointed toward the child in the educational enterprise.

He failed. If only he worked harder, if only he tried more, or, perhaps, if only he came from better stock, but it was essentially his fault ... he failed. Then for a variety of reasons the finger subtly changed and moved away from the child and pointed toward the family. The family wasn't providing the appropriate learning supports, the appropriate kind of motivations, and perhaps, the appropriate genes. It didn't stay at the family very long because parents don't like to be pointed at in that regard, and there were other good reasons why the finger moved on. Rather it moved on very clearly to the school. The school is now the culpable party. The enterprise is interpreted as failing, and the blame is being put on the schools. I sense now that the finger is moving again very subtly, and it is moving directly toward social science. That is, the schools, the families, the children, and the politicians are saying: "Okay, the enterprise is failing, schools are not doing a good job, and we admit it. Families, for whatever reasons, are not doing a good job, and we admit it. The kids are still failing. The problem is to teach the kids, not to fail them. Tell us how to do it better." The finger is pointing at social scientists, and I think it is not so much a sign of culpability as it is a sign of a weakness. It is pointing at us because we are the weakest party in this enterprise right now. How can we become stronger? How can we organize or engage in educational research, development, and evaluation in a way that we will have some impact and power? That's a question I don't know how to answer.

OLSON: We have been collecting reams of data, particularly in Title I, on the context of the educational enterprise, yet we have not been able to relate it to what few outcome measures we have. I think what you are talking about in terms of context is important but I'd like to hear more, especially for the developers. Would you care to elaborate?

SIEBER: I'm not sure that I understand the question; are you saying that you have been unable to relate your inputs to outcomes?

OLSON: We collect data on context, socioeconomic status of the youngster and all the rest. We also collect some outcome measures. In Title I, he's disadvantaged in the first place and if we find that he isn't doing any better, we might be inclined to say, "Well, what can you expect," and the like. Can't we do better than that in terms of relating outcomes to context, and then by trying to do something about the context?

SIEBER: I don't have any answer, I'm sorry. I don't know what strategy it would be. I feel that I could find some answers as a kind of qualitative observer as well as a quantitative researcher using a variety of techniques, not just experimental design but also elaborate questionnaire and statistical controls, depth interviews, etc. That is, by studying that complicated program with every kind of technique that is at our command as social scientists and somehow coming up with results that this variable had an effect, that a second variable did not have an effect, that a third variable seemed to be a hindrance, and so forth. What we need is a much more formal kind of strategy or scheme for this, to produce the best evaluation possible right now.

OLSON: You have very little to offer the curriculum developer, it seems to me. He still is at a lost as to what to do about change.

MESSICK: You can look at differential effectiveness of change within context although there are some arguments about this. There is an approach called "educational performance indicators" that operates at the program or school level and that takes into account the prediction of outcome by contextual and hard-to-change characteristics in the situation, like socioeconomic status. You find that you can get a nice regression line with schools varying around the line. In general, schools located in high socioeconomic communities (with better resources and other things) produce better outcome results than schools located in low socioeconomic communities. To evaluate the differential effectiveness within contextual level, you would ask, "Are there schools at the low end, admittedly not producing very big outcomes, that are producing better outcomes than you would predict given the resources they had available and given the input characteristics of the children?" You might find a school that is doing a very good job in the sense of being far above the regression line down at its low end whereas another school at the upper level (given all the good resources and the "desirable" student input characteristics) might be far below the regression line. If we are concerned with evaluating schools, we might say this quite rich school, using rich in a very global integrated summary sense, is not doing as good a job as this poor school because one is above the regression line at the low end and the other is below the regression line at the upper end. Then you could ask what are the correlates of the deviation from that regression line, and this would give you a handle on procedures that could be introduced into other schools to improve effectiveness within their own particular context.

This performance indicator approach has been criticized extensively by people who are concerned about the poverty community and ethnic minority groups by saying that we might become smugly satisfied with fixing it so that all the schools down at the low end go up above the regression line although they are still very low, in an absolute sense, on outcome measures. They say, "That is not enough. Ultimately we don't care whether we're above the regression line or not; we care about the absolute value of the outcome. We want the regression line changed and we would like to see it flat and very high."

OLSON: What you just said seems to me to be a very important ingredient in this consortium concept. That is, if we can find evaluators who can supply what you're talking about, then the evaluators are going to be able to supply information for so-called developers who then can do something rather than just operating in a vacuum. I know that the state of the art is pretty shallow at this point, but it is the part that I think we need to concentrate on, otherwise we have nothing for the developer to do with what the evaluator is supplying.

MESSICK: One of the problems is the pluralistic nature of the clientele that we are dealing with, but as long as we can agree on certain outcomes as being very valuable and very positive then I think that strategy is a good one. But there are large segments of the educational community that are saying that they do not agree with the outcomes. How do we respond to that? Do we say that it is okay, that in a pluralistic society any sub-group within the larger culture has a right to determine its own destiny, its own values, and its own goals ... even though those goals might be counter-productive in the sense that they are not supportive of ultimate survival in the larger society.

LUMSDAINE: Well, one thing that you can do, of course, is try to "present them with the facts," to use that trite expression. That is, you can at least try to show what outcomes eventuate from particular contexts. This includes both intended and, to the extent that you have the wit to anticipate them, the unintended outcomes resulting from a particular educational procedure. The assumption is made that the procedure has some element of reproducibility in that the evaluator can say, "If you do this, you will get that result." Such information is useful even if you can't fully agree on how desirable it is to achieve this outcome or that outcome. I think that I would agree fully with what I understood you to be saying. This use of evaluative data is of great value; at least you know what is happening as a result of what you're doing, and you are, at least potentially, in a position to do better.

ROBINSON: Seemingly implicit in the discussion that preceded was the notion that the relevant variables, and the measures that define the dimensions of these variables, are known. As a developer with some playing around with evaluation, I have a very different feeling about that.

MESSICK: On the contrary, my remarks certainly assumed that that was the problem. If we knew the nature of the variables that were operating, the dimensions of the problem, then development follows relatively routinely from that understanding.

ROBINSON: Except that in your model, there seemed to be one thing missing that is quite critical. That element is evaluation of the outcome of the development group in terms of the design which was handed to them. It seems to me that in a lot of cases, both in development of measurement devices and also development of curriculum, the specifications are sometimes quite well done, but the carrying-out of those specifications by the development group falls very short. Still, no one goes back to systematically examine the process of development.

MESSICK: I was assuming that development would always be evaluated and redirected appropriately. Again, let's take the development of tests as an example. Elaborate machinery has been developed that essentially asks, in terms of empirical evidence, do the test items that were developed to meet the specifications in fact meet the specifications?

ROBINSON: There's a heck of a lot of judgment that enters in rather than just empirical evidence.

MESSICK: Well, in the development of a particular test even with highly-defined specifications, you typically develop twice as many or three times as many items as you ultimately plan to use; you throw away items that are found wanting.

LUMSDAINE: Wanting in terms of what criterion, Sam?

MESSICK: Wanting in this sense; let's take a simple example. Assume that you have a specification grid that indicates that you'd like to have an achievement test that would cover a subject matter area, say the typical curriculum in American History. But there are certain processes that you also want to tap, and you would like to have coverage of all the processes over all the content topics. You might decide to write 10 items for each cell. Well, that's a hypothesis ... you are hypothesizing that the items that you wrote indeed cluster together empirically. Next you get information about the intercorrelations of the items, item-total correlations, and things of that type that enable you to evaluate whether or not you've done a good job in covering that domain. At the same time (and this I think is very important), you also can get information that might indicate that the domain as specified was incorrect. Frequently in looking back from the empirical data to the original specifications, we notice that what was believed to be a single cell is really two dimensions. You then can go back to the specifiers, and they might agree that an important element of the situation was neglected. Therefore, they change the specifications on the basis of empirical results. If this were done routinely and continually, it would lead to a theory of achievement in each subject matter domain. I am not suggesting that we do that, however. Certainly we at ETS don't nor does anybody else; as a result, there is no theory of achievement in any subject matter domain.

ROBINSON: When we go beyond that rather simple kind of problem, it gets even worse because of the frequently-occurring hiatus between a cluster of items, even supported by empirical data, and the specifications. Judgment seems to enter in the resolution of such discrepancies. For example, I can say that the items I have developed really get at this idea, and someone else can reply, "the heck they do." There's no resolving criterion except the way each of us feels.

MESSICK: At one simpler level below that, the question becomes (even with judgment), whether or not all 10 items developed for a certain cell get at it to the same degree. Empirically, we can decide on each item. Whether it's an important thing to get at is still a judgment but we can say that you have written only 2 items that are any good, not 10.

ROBINSON: It would seem, then, that a major part of the need in evaluation and development is for a tremendous input into devising and searching for variables, new variables, and an input directed toward perfecting measurement devices.

MESSICK: You're right. I think that is a very critical issue. The really difficult thing we have to face is understanding the dimensions that are operating in the problem area. What are the variables that are important to assess? People consider that question to be in the research domain, and research is a frill these days although, as I noted, research is much cheaper to support than action programs.

Also, in terms of funding, we operate in educational development on a very different scale than in most other areas. We really have very little money. I was present when someone in Washington asked, "How come educational researchers can't tell us what to do?" (This happened to be in the early childhood area.) The question was: "How come with all the research that has been done in child development and early education, you don't know what to do? Why, that's terrible!" In the ensuing discussion, it was pointed out by a child psychologist that the questioner really didn't understand the dimensions of the funding problem. It was pointed out that the operating budget for Project Headstart during its first year (which was $300,000,000) would have paid for all of the child development research all over the world for the past 40 years, including Jean Piaget's salary. Now if you translate that into other units like aircraft carriers, you can see we are really talking about a pittance.

Even in areas that are well-researched, there are important development problems. Take, for example, certain kinds of verbal aptitudes. We have 50 years of research behind us on the nature of the dimensions that are operating. Where you have prior research and evaluation, then the developmental problem became a really critical one. It involves how to get impeccably-developed instruments that have optimal properties. If you're working in an area, say self-concept, where the dimensions are not well-defined, then the development problem must be intimately related with both the research problem and the evaluation problem. At the training level, you are not going to train people to be developers in these narrow areas, that is, to be developers in self-concept or developers in verbal aptitude. Since development is intimately and inextricably intertwined with problems of evaluation and research, the training of people to be developers must occur in those contexts.

LUMSDAINE: Sam, awhile ago we found or I found, that we were in agreement when I thought we were in disagreement. Now I think we are in disagreement when I thought we were in agreement. I agree with your statement that in those areas where the necessary research has been done, then it's primarily a question of development. The only thing is, I don't think there is any area like that.

MESSICK: Okay.

LUMSDAINE: I think that in any area we consider, we have to check whatever presuppositions we have from background research with the sort of "proof-of-the-pudding" evaluative data that we collect.

The other point that I would like to make is in relativistic terms. It doesn't seem to be a question of how much percentage-wise should go into development or evaluation, but rather that more should go into those enterprises that intimately combine development and evaluation so that we're not talking about one or the other, but rather the combination of the two into a single activity.

GUBLER: Organizationally, this has raised some rather interesting questions in terms of objectivity. If you intertwine evaluation and development to such an extent that they are no longer separable, then presumably the same people in the organization are determining them both.

LUMSDAINE: Very good point. I think that this again raises the question of two different kinds of evaluation, formative and summative. I personally feel that the most mileage is to be gained out of formative evaluation in which the evaluative data feed right back into an iterative development-improvement cycle. But also at some point you do have to have, by some means, a non-incestuous kind of evaluation in which someone from the outside comes in and says, "Okay, you feel that you have done your best using evaluative data to make changes. Now we are going to make an independent assessment of how well you have done."

GUBLER: I'm going to try another idea because I can conceive of a situation where you develop an expertise to take the data accumulated by the evaluator, transfer it into layman's language, and, using the public media, create a different kind of evaluation mechanism. Maybe that is what we're lacking.

Now I'm back to one of my initial statements. Have we gone so far astray in terms of our jargon and in terms of our data-gathering processes that we have somehow left the animal (that's suppose to benefit from it all) by the wayside? I just raise that question.

LUMSDAINE: I take it that that's a rhetorical question.

GUBLER: It may be, but it is a practical question.

MESSICK: Maybe we left the animal that is supposed to decide who's to benefit from it by the wayside.

GUBLER: But he can't decide unless he has some of that data.

MESSICK: We have to leave in a moment, but I would like to emphasize that I believe where Art ended up on this last issue was in support of my conclusion. He didn't particularly like the premise from which I arrived at that conclusion.

LUMSDAINE: The retort courteous to that is, "You keep off my premises."

MESSICK: One other point. I'm essentially a personality researcher and I know that area fairly well. To do personality research, you have to have measures of the variables that you are trying to interrelate and understand. So you develop measures in terms of your best understanding, you evaluate the adequacy of those measures, you do further research which leads you to re-conceptualize the variables, you then redevelop the measures which leads you to re-evaluate them, and you go through an iterative cycle in this way. The original motivating force was research-oriented, that is, trying to understand the nature of the variables in the system. The process that you went through taps development, evaluation, and research, and it is iterative. If it's a continuous iterative process for the developer or the evaluator, then I don't see that it matters where you enter the iteration, as long as you replicate it several times. But if you enter the iteration, say, at the development stage, and you develop and that's all, then I think inevitably, regardless of the area, that product is going to be found wanting.

I also don't believe that it's a good idea to enter and engage in the iteration only once, no matter where you start. That is, if you start with a presumed understanding of the research domain and say that research leads to development which leads to evaluation, and you find that the result is good and you bless it ... almost certainly it's inadequate. Research should lead to development to evaluation to research to development to evaluation ... we should recognize the iterative nature of the enterprise, and we should recognize that the skills and demands of the situation are intimately interrelated. We should recognize this at all levels, including the initial graduate training for this effort. As a researcher, I seem to keep emphasizing that research is a critical part of this; I see others as part of the research effort. If I were an evaluator, I would probably see the others as part of the evaluation effort; and if a developer, I would see others as part of the development effort. They're really all part of the same repetitive process, a cyclical process, a process which absolutely demands feedback. Any model that conceptualizes this domain as a linear one, of one thing leading to another, is absolutely incorrect. I will be very dogmatic on that point. Without feedback, we cannot proceed properly in this domain.

GUBLER: I presume what I'm saying is that the research function is a more delimited function than the evaluation function, and they move from different premises. The research function to a large extent is imbedded in, although not exclusively, quantitative analysis.

MESSICK: I would say research is imbedded in conceptual analysis. If you say conceptual analysis, then it makes everything else a part of it. I would start with research being the conceptual process, while the quantitative techniques of evaluation and development would be part of that process. Always, however, the critical aspect of this is the conceptual process. The decisions are going to be made conceptually in terms of values, so I would put primacy on the research side, but then again, I am a researcher.

LUMSDAINE: Sam, I think that's a good note on which to conclude.