REPORTS OF THE THREE WORKING PANEL CHAIRPERSONS (CONTINUED)

II. WORKING PANEL II: IMPROVING THE DEFINITION OF EVALUATION CRITERIA

CHAIRWOMAN: MARCIA GUTTENTAG, Director,
Social Development Project
Harvard Graduate School
of Education

MS. CHELIMSKY:

Let's hold any questions and go right on since there is not much time left. Marcia, would you like to tell us what Panel II found?

MS. GUTTENTAG:

Our task was to discuss improving the definition of evaluation criteria. As someone left our meeting at the very end, he said to me, "Is it really possible to disagree with everything that everyone else has said here?"

With that as a caveat, whatever I present is necessarily distorted and shaped to make it sound as though there is a little consensus around what was said. I have six points to make which summarize our discussion, then a couple of conclusions, and a pessimistic and an optimistic note on which to end.

Point Number 1. We began by discussing effectiveness and efficiency criteria and decided early on that these were only two among many and that it was important not to use them as the sole criteria, this for several reasons.

First, that there are an enormous number of different means and different ways of operationalizing each of these constructs

and that, because of this, any single choice in operationalization is bound to exclude others.

Second, that they often are premature specifications as categories. That deciding that one is going to use them as the evaluative criteria makes certain presumptions about what is being evaluated which may not be correct assumptions either at that time or at any time.

The second point. That led us to a discussion of the realities of evaluative criteria, which we think are partially, first, that there are many different audiences, different users, multiple perspectives which each of these audiences have and therefore multiple and different criteria. That the values and criteria of these audiences must be specified in one way as one of the basic ways of defining evaluative criteria. That is one of the first jobs one has.

The third point. We then entered a discussion of the difference between criteria and measurement. If I can summarize that, we felt this a very important distinction because criteria are never subsumed by any single form of measurement. Criteria are the standards or objectives--that is, they are much more abstract than any single set of measures. The measures are the forms of information which are related back to the criteria.

Fourth point. We then turned from this relatively abstract discussion to a discussion of what are the concrete criteria that are now important in decision-making in various agencies--the agencies represented in the room. I am going to mention three of these.

One criterion which came up was compliance. That was the criterion being used by the Internal Revenue Service.

Another criterion which was discussed--and this one seemed to be extremely general across agencies--was influencing other organizations.

A third criterion was institutionalization of a program--that is, will someone else pick up the tab.

Let me reindicate what we are talking about here when we talk about criteria. These are the criteria that are used to determine what decisions will be made about. That is decisions are made on the basis of answers to these criteria. It is clear having presented these criteria to you, that most evaluators are not in the business of providing information relevant to them of doing measurement that is relevant to them. We thought that was quite interesting.

Someone in the group suggested that perhaps one other transcendent criterion that all agencies have is some aspect of cost-benefit analysis, or how much things cost.

Fifth, we then turned to a discussion of how to avoid what was characterized as Type 1 errors. That is, the consecutive shaping of the criteria in terms of what looked as though the measures would turn out to show to be successful.[40] This is the old problem of looking under the streetlight because the light is better there for the keys that you lost down the street where it was dark.

In other words, this is a danger raised about fitting one's criteria to one's successes and successively pruning along the way so that the measurement that was finally decided upon would be a very carefully selected set or single instance of gems rather than stones.

---

[40] See page 277, footnote 36, above.

The sixth and final point is that I think we all agree that evaluation should only be done where information will be used in decision-making, and that it was pointless to spend evaluation resources to conduct evaluations where the information would not be used in decision-making.

Trying to end on a hopeful note, we did that the following way. First, one of our group suggested that there are lots of simple questions which are essentially descriptive that can be answered and are being answered all the time. We are doing better at that.

Second, that a great many decisions require simple, descriptive information about what is happening and what is related to what. We are equipped to answer such things. It's only farther down the pike that we get to questions of why it is happening; and although we are very interested in those issues and those are the complex and policy-related issues, those are not the questions that are being asked.

We ended on a very opportunistic note when Kenneth R. Feinberg said that for the evaluator king who can come along and discover what programs will actually reduce the crime rate, the presidency is waiting. What I have given you is a list, and you might be interested in what we think is important and not important on that list. Perhaps I should leave that to the questions and to let other people in the groups--in the group that we had, answer.

MR. BENINGTON:

While people think of serious questions, I have a comment. I now see the statistician's view of the Constitution. And that is that the Executive makes Type 1 errors and the Legislative, Type 2.

MS. CHELIMSKY:

Are there any questions, serious or not?

PARTICIPANT:

I am Vickie Jaycox, the National Institute. My major complaint from this discussion goes back to what I feel is sort of a cop-out of evaluators at this point. Federal agencies, regardless of who the user of the evaluation is, at some point have to answer to whether or not they have had any effect on what the legislation was formulated for--basic questions of whether or not they are going to get refunded, related to whether their programs changed anything in the world. So when you get down to the question of whether Joe Shmo wants to refund his program, he has different questions to ask. When you are asking whether something changed in the world, then you are into a different kind of evaluation. Now, what everybody has been talking about is a very simple, straightforward, user-oriented evaluation. But there has been really no discussion of the role of really basic evaluation research, asking what is the effect of programs. I think it's something that was missed. I'd like some kind of comments on whether we are ever going to get back to real basic effectiveness evaluations, on whether we learned something conclusive from the evaluation. Does that make any sense?

MS. GUTTENTAG:

Absolutely. I hope I am free to give a personal opinion. Must I keep representing the panel?

Of course that is the question--does what we are doing matter in any way? What are the effects of what we are doing? I think that is the key issue in evaluation.

I personally though have been quite biased in reading evaluations by looking at the evaluation methodologies that have been used to answer that question. The methodologies themselves have often been inappropriate because of the assumptions that they have made, either about what is happening in the world, or about the statistical properties of what is happening in the world such that certain methods could be used. So, coming from that critical stance, I am always extremely concerned about what I call premature effectiveness evaluation. That is, it seems to me that more untruths have been told in the attempt to try to say what the effectiveness of a program is than the reverse. That is, I think we have been on safer ground in looking at a variety of different criteria and in keying evaluations to the criteria that decision-makers have so that the information that is produced is always in terms of the decisions that have to be made as a program develops. That is terribly abstract; I suppose that is why you get something of a bias in what I have said.

MS. JAYCOX:

I feel that it's because it's so difficult that we say, "Well, we don't want to do that anyway."

MS. GUTTENTAG:

My opinion is that we must provide decision-makers with information that they want.

MS. JAYCOX:

At a higher level, that is a very demanding kind of information. Did you reduce crime this year? We can't tell. We don't know.

MS. GUTTENTAG:

That's right. Well, you see, I think we are so much better off saying, We don't know, than, No, the things we did didn't help.

QUESTION FROM FLOOR:

Is it that you say,we don't know,or that you illuminate the number of things that affect the crime rate beyond the narrow things you measure?

MS. GUTTENTAG:

That is certainly one of the very useful ways of answering that question.

PARTICIPANT:

Walter Bergman, IRS. I share the same concern as that expressed by Ms. Jaycox. I think the answer is really long-term research as opposed to what I have only learned in very recent months or the last two years to know by the name of evaluation. Because it's a term we never even used. This takes more than answering a single administrator's politically motivated, generally immediate whim. I think it transcends administration. I think it transcends a single manager's interests. I don't think these answers can be gotten easily--I have to keep talking about IRS because it is something I know about. In our particular instance, we started in 1962 with our taxpayer compliance measurement program. We are trying to find out not what our body count is, but we are trying to find out whether we are doing anything to affect the public out there in terms of their behavior, their compliance. And what is happening to it. This does require a serious experimental design. We have had to develop panels. Unfortunately this means the same person gets audited twice in a row. We are trying to find out whether or not the fact that we audited him the first time made any difference in his behavior the second time. Fortunately, I would hope that our process is not considered destructive testing.

MS. GUTTENTAG:

Do you have informed consent to that?

QUESTION FROM FLOOR:

Informed by whom?

QUESTION FROM FLOOR:

Is it random assignment?

MR. BERGMAN:

Yes, it is random. It's random within random. We do get protests at times, I assure you. But somehow, we have been able to convince our constituency that this is necessary in order to maintain a voluntary compliance system with the tax system.

My only argument really is that I think we have to differentiate between long-run research, which will give us some insight into the real hard answers -- the final outputs that I mentioned before, versus some of the shorter-run evaluations. I think evaluations are wonderful for efficiency kinds of measurements. We do a lot of those, too.

MR. EWING:

Could I comment on that? I'd like to say that it seems to me that if you got at some juncture a willing ear on the part of a program manager or agency head -- whatever he might be -- decision-maker, somebody who is willing to talk about what his goals and objectives may be, one of the aspects of that situation is that you have got an interactive kind of discussion going, hopefully, in which he says what his objectives are; and you tell him what you can give him in the short run (if he wants something in the short run) and what you can't. You also tell him what can be measured currently

and what can't be or how well it can be. It seems to me that in keeping with our notion of a deal or a bargain, what you are hopefully able to work out is some kind of agreement that there are some things that can indeed be answered today or tomorrow or Friday or next week, and other things that will take a year and maybe will result in nothing much more than a disciplined judgment. Some other things require systematic analysis. Some things require monitoring only. Other things require evaluation of a fairly well-disciplined sort, and some other things can only be answered through long-term research.

Hopefully, you can work out therefore a kind of a mix of strategies which, combined, will begin not only to answer the perhaps politically motivated, short-term administrator's question, but also begin to serve the function of accumulating knowledge, putting building blocks in place and beginning to build a body of knowledge from which much more sensible judgments and decisions can be made.

To respond to an earlier question about where the agencies are that have done this, I don't know of any that have done it; but let me just say on behalf of one that has been much criticized, both here and elsewhere, that LEAA has at least put together an evaluation program which includes evaluation of its discretionary funds which are program evaluations, many of which are very clumsy and awkward kinds of evaluations. But it is also developing a program in the development of better methods for measuring and is also working on developing instructions to states and local governments on how to do simple evaluations and more complex ones and is also doing some long-range kinds of things including some things that started a couple of years ago -- for us it's long-range. For most Federal agencies it is. They are going to last another three or four years.

I think it is probably true that a great many agencies are working in that vein trying to put together a mix of strategies.

PARTICIPANT:

I'd like to make an observation, at least. You know, really, most of us who are here representing an agency are here because, to some varying degree, that agency is supporting an evaluation effort. In varying degrees, we are or have recognition within the agency. It seems to me one of the things we have on occasion in the conference overlooked is that we are ourselves in most cases managers who have evaluation as a product. We are responsible therefore as managers to really do a great many of the things that we are ourselves in turn talking about trying to get managers to do.

It seems on occasion we have to talk about negotiation. We have to negotiate our own products, all right. We have to sell those products as evaluation, and I think what we have had represented here also on occasion are a multitude of different management styles as well as evaluative styles. Some have reflected management styles that have worked. There are those, for instance, in education, who have said, you know, we have had an office that has been able to accomplish a certain sale of our product.

In other words, we are ourselves managers, and it seems we are mixing on occasion a management question with a product question. That has been a part of our conflict here.

PARTICIPANT:

I'd like to solicit your comments on how you feel about the same thing. I have gotten the feeling that there is sort of a projective need on the part of decision-makers, as compared to the

overall feeling of a retrospective view in evaluation. The decision-makers must take not only a retrospective view of what worked and what didn't work (and possibly why and how) but must also address the "what if" question. My decisions relate not so much to what has passed, but what is in the future: if I have options, or if I can identify alternative options, I need to have some means--sometimes it's the seat of the pants, it's intuitive, it's mathematical, call it what you will. All of those. But how do I convert what happened in that case, that set of cases, into the decisions I have to make about what will happen or what is likely to happen? How do I convert the retrospective into the projective?

We seem to have been focusing on what happened, and I don't know how we are going to get into crystal balls, map modeling, seat of the pants, how we are going to put these things together. But most of the decisions are not retrospective. They are projective. I'd like to get your feeling on how we convert an evaluation of a project that is on-going or that happened into projective tools that are credible?

MS. CHELIMSKY:

I think one of the big problems we have is that an evaluation's findings are often not generalizable (because of problems in the design, because of problems in the data, because of a million other reasons) even for the period in which they are derived. So that, you know, if they aren't even generalizable beyond the population studied to begin with, it's difficult to have confidence in their generalizability to unknown future situations.

PARTICIPANT:

As I see it, the basic problem of a manager is to generalize. And the point is, he either generalizes to different individuals, different programs, or to the same one continuing or changing. His is

inherently an objective task. I don't think we are facing the fact that our view has been essentially retrospective, but his view is essentially projective.

MR. EWING:

Let me comment briefly. It seems to me the usual scenario in most agencies is that people whose background is in research become evaluators or become the managers of evaluation which is contracted out. They get products which then get sent in nice neat packages to administrators of agencies, and the administrators don't read them because they are too thick or because they are untimely or because they simply have no training or background themselves which permits them to make head or tails of what is given them. Most administrators for some reason -- I'm sure there are reasons -- are not themselves trained in research or have any experience with research.

One of the things that is missing is a bridging function. We talked about that some in our panel. A function that involves somebody who understands enough about research to understand what it is the evaluation results say, but that same person has to understand enough about the needs of management to assure that he can take management's needs and make sense of them in terms of the evaluation results. That is a rare kind of person who can do that. It's a function that gets performed, I think, very seldom. It's one that I think most agencies have a great deal of trouble with, but it's not an impossible thing to do if somebody is assigned to do it who has some common sense. One of the troubles with it is that it hasn't been recognized well enough as a discrete function which needs to be performed and which is not typically well performed by a researcher or by a manager by himself.

Our administrator, for example, is fairly interested in evaluation results, but tends to be put off by them the more they are put in terms which he regards as research gobbledygook. That I think is a serious problem.

Related to that is a comment that was made in our panel which is that a great many people seem to make evaluation a very pretentious kind of thing. That is, more pretentious than it needs to be or deserves to be. If it were stated more modestly, it would not only be better understood, but more in keeping with the modesty of the findings. That might also help.

MS. GUTTENTAG:

There are of course models of inference which make it possible to take a prospective look. They are available.

END