X. THE NATIONAL EVALUATION PROGRAM: KNOWLEDGE SYNTHESIS

JOE N. NAY, Senior Research Associate,
The Urban Institute

MR. GRANDY:

Joe Nay is going to present his perspective on knowledge synthesis. Joe is currently at the Urban Institute. He is an engineer by training, a graduate of a joint program between the Electrical Engineering Department and the Sloane School of Management at MIT. He has done quite a bit of work primarily with interdisciplinary teams to alter the operations and improve the effectiveness of large organizations, both inside and outside of Government. His experience covers management problems, policy research, practical problems of implementation and also evaluation. Joe, it's a pleasure to welcome you.

MR. NAY:

After listening to everyone else yesterday, I reworked my talk last night. I don't know if I have done a good or a bad job yet; but I'd like to start with something that happened to a friend of mine a few years back, which, I think, puts some of the things you heard yesterday in perspective.

This person decided to do a series of interviews with high-level analysts and high-level policy people in a series of departments in the Federal Government. He collected a lot of names from many of us. He interviewed a lot of the analysts and I was very interested in how it came out because I had been close to the past work of several of those analysts. A lot of their work had had effects that I knew about.

Some of the effects were very positive. Some of the effects (I thought) destroyed things that I was very fond of. I had a lot of mixed feelings both about how the effects of their work had come out and how all of this would come out in the interviews.

I think that both of us were astounded when he came back with the first round of interviews. Almost universally, people in these analysis and staff groups had told him that they hadn't had any effect at all. I plowed through some of the interviews myself with him. I even found that some people whose effects I knew of (because I had been working with line management people at the time the effects of their work took place) had said, "The most frustrating thing about my three-year tour was that I didn't have any effect at all." How could they say that?

I sometimes think that people in staff groups and a lot of evaluators and analysts, in particular, have a vision in their head that is left over from "Executive Suite."[32] That serial has done more harm to management than anything else that ever happened. It left people with visions of the big meeting where decisions are made. Everybody has a cigar, and they say, "What shall we do?" The analyst reads off his numbers, and they say, "That's it. That's it. That is what we are going to do!" Few analysts ever actually find themselves in such a meeting; perhaps that is why they think that their work has no effect. If you look upon evaluation as gathering information to have an effect on an organization or upon the decision-makers in that organization, however, I think that you have to look very carefully at the sort of ripple effects that each effort has.

---

[32]Editor's Note: "Executive Suite" was a movie, genre soap opera, serialized on television during the fall of 1976.

In one sense, I think a lot of those analysts were right. They had often gone to meetings and taken their papers with them. They said, "This is what we ought to do!" And the decision-maker didn't do exactly what they said. But in some particular cases that I knew about where my friend found the interviewee still saying, "None of my stuff had any effect," I knew that in many cases it had had widespread effect, either by altering some course of action, or preventing another one, or by really sealing a choice that people hadn't quite made up their minds to make.

So I think that even the idea of effect is more in line with what Donald Elisburg said last night.[33] Whether something has effect or not depends upon what different people will accept as proof and how their actions are influenced, or bounded, by information that they believe.

The National Evaluation Program at LEAA is partly a knowledge synthesis program. It's broken into a Phase I study which is a synthesis and assessment study and larger Phase II evaluation studies. I'll talk a little bit about how that came about.

A Phase I study is really a synthesis of the information that is available. We could talk for hours about what I think is necessary and unnecessary to do knowledge synthesis, but I want all the Phase I grantees to stay in the room so I'm not going to give that talk. This way, the Phase I grantees won't have heard this entire talk already.

The important thing about the NEP (after hearing yesterday's high-level people from agencies around town) is that it is something that has been carried out. A lot of information has been gathered together. A lot of knowledge files have been built. It is kind of interesting to see how that worked. Our role is as technical advisor,

---

[33] See page 201 above.

and we are doing a case study of how it all happened over the last two or three years and how we think it all came out. We are also giving intermediate advisories along the way of things we think ought to be changed.

The present emphasis on oversight is one of the factors that is leading to the development of these syntheses programs in several agencies right now. And acceptance of the results hinges in part on degrees of proof. English is a funny language. There are two definitions of "oversight." The first one is supervision, superintendency, inspection, charge, care, management and control. A lot of people forget that there is also a second definition of oversight that is used every day, which is the fact of passing over without seeing, omission or failure to see or notice, inadvertence.

I want to talk today about a real life attempt by an agency to convert what a lot of people thought was a case of the latter definition to a case of the former definition, the National Evaluation Program.

When I used to try to teach people about evaluation in Government programs, I always required that they look at a program and find out some very simple things at the start. I used to keep pounding, "Go out and look and see if it exists." People say, "Evaluators haven't done anything." But there are hundreds of programs around the country that never were implemented in anything near the shape in which they were envisioned. And without evaluators, no one would ever have known this in many cases. I think the evaluators have pointed that out, and I think that is a valuable function. So the first question about a program is, Does it exist? and the second question is, What is it? What process is in operation? What is it that exists? What outcomes are produced (you have heard all these before in any evaluation paper that you have read) and what impact do they have?

We can't do any less for the NEP. There will be a case study out in May where we will try to answer those questions for the first two-and-a-half years of the program. But we can answer the question now (sort of from the laboratory to you) although we may have to reverse ourselves later. We can say, Does the NEP exist? Yes. What process and operation? We can't tell you all about it today in a half-an-hour, but we have it pretty well documented. What outcomes have been produced so far? Nineteen studies have been produced, and there are eight more underway. There will be another batch next year. What impacts do they have? Some of those impacts are being captured through surveys and interviews. Others won't be.

For a number of years, as a couple of people have remarked, the bulk of LEAA money went into the block grant program. The block grant program was originally, by design, a case of the second type of oversight. At one time it was characterized as "leaving the money on a stump and letting someone come and get it," the way people used to buy moonshine. This was a result of an argument about whether local initiatives or national categorical programs were better; and for a long time, LEAA had this block grant program. There were tens of thousands of grants out there, hundreds of most any kind that you could name that were commonly known. They were locally determined, and most of their evaluation, if it was done at all, was done locally. Most of the national evaluation effort was made against the discretionary money, on that part of the money that national LEAA controlled.

The 1973 Act required oversight in evaluation. If you can picture what happened, you go along for a number of years. You give away your money. People make grants with it for things that they think are good. Suddenly Congress says, "You don't know what they are doing. You don't know how it's working out. We want some oversight information about this."

Most people suggested that four or five big evaluations be done immediately, that large, long-term evaluations with clear assumptions be put in the field. The problem was that when all the internal suggestions were produced of what should be evaluated, there were (on the last list that I could find when I was preparing this talk) 122 topic areas that people had suggested as needing one of these five costly evaluations.

Many groups in Government have been faced with similar problems, and I think many groups have called in the universities and selected five topics and begun large-scale evaluations. Some of these have efforts worked out; but, as you heard yesterday, an awful lot of them have run aground. They have come back with findings about the nature of what is out there. What was being done in the field has been different than everybody thought. The measurements selected in advance by the agency and the evaluation grantee haven't exactly fitted the programs to be measured. There has been controversy about the results.

LEAA did, we thought, a clever thing. They convened a task force whose director is in this room and settled upon a strategy of trying to milk knowledge in sequential steps from those locally-determined block grants in order to go at it in stages and try to build some information files. A little over two years ago, they came to us and said, "We want to try one of your approaches of buying knowledge in sequential stages." That is always a pretty good thing. It makes you feel good if they say they want to try one of your approaches. The bad part was they wanted us to help. After a lot of hassling over the ground rules, we agreed to serve as technical advisors and to do a case study of what happened.

In the face of all of the same pressures and problems that were outlined to you so gloomily yesterday, of pressures from up above,

pressures to hide results, vagueness of objectives, certainly a lack of consistency in many of the programs, enormous gaps between theory and practice, the National Evaluation Program has come into being. It has produced the 19 Phase I studies that are complete and has 8 more underway. Despite the problems that you heard about from executives from half of the Federal Government yesterday, the full studies are available. You can get them. You can check them out of the library or you can get them on Microfiche. Some are better than others. You can get them all. Summaries of all are being distributed.

The summaries which are written by the grantees are nationally distributed. Some demonstrable impacts have already occurred, and we are following up with surveys and interviews to try to check out some more. Every study has been preliminarily rated, both whether it's the kind of thing we thought we were buying with Phase I work descriptions, and on what we think the apparent usefulness of it is. The program has been kept stable long enough that we are beginning to have a good idea of what some of its strengths and weaknesses are. Changes are now being made to improve some of the problems that have cropped up.

In May, as I said, the case study will be available; and you will be able to see what we think about the whole process.

In light of what you heard yesterday from various officials who told you why something like this cannot be done, it's hard to understand how this could have happened. So I've revised my talk on knowledge synthesis to try to outline for you here today the key things that I think allowed it to happen. I have five here. (There may be a different five in the report.) They are:

- Simplistic thinking
- Stubbornness
- A detailed approach
- Pressure to follow it
- A single person in charge

Let's see, simplistic thinking and stubbornness. I think people sort of outlined some simple things to do and they stuck with them for a year or two, an underlying concept or two that didn't get modified until the agency could begin to see how they worked. Unusual, but it happened. Two more key factors were the work description (i.e., a detailed approach) and pressure to follow it. I think the fact that a single person was responsible for it (Dick Barnes[34] who is back there in the corner and ought to be up here speaking) is major. He has stuck with this thing for two-and-a-half years. He has been responsible for it, and he has been the focal point for it. He has gotten encouragement and occasional discouragement from the heads of his agency and other people in his agency. He is still on the program. I think his strong determination to do these obvious things--read the proposals, look at the concept papers, talk to the grantees, try to get people to modify their approach a little bit so they come out a little better--has been a key factor.

One of the simplistic ideas was that too little was known about what was actually happening in many topic areas to really begin full-scale evaluation. This led to the idea of a Phase I, Phase II exploration. I will not talk about Phase II today.

Phase I is really a form of evaluability assessment, and we will talk a lot about the nature of what we think evaluability assessment is.

---

[34] Editor's Note: Head of the National Evaluation Program at the National Institute of Law Enforcement and Criminal Justice.

Phase II is a larger, longer evaluation where one appears warranted, and after you know enough about the area to better begin to scope one.

People talked a lot yesterday about the dangers in the evaluator's job. There are a lot of dangers in the evaluator's job, and I believe Jim Stockdill noted that the evaluators may often be the only persons who are looking at both the rhetorical charters and the operating activities.[35] From the standpoint of an organization trying to implement programs, you don't want to ever sell that activity short because questions about performance come from those rhetorical charters in many cases. The measurements that will have to be taken if an evaluator does the measurements himself will always be out where the activities are. When we talk about evaluability assessments, we are trying to assess that gap and bring the rhetoric and the activities closer together before buying major evaluations.

Again, you have heard my stories before. There is a favorite quote of mine in one of Shakespeare's plays that goes something like this. One fellow says, "I can call dragons from the misty deep." And the other replies, "So can I and so can any man; but the question is, when you call them, will they come?"

Now, various private and public groups have been busy calling those dragons from the deep in the form of policies and even programs to solve problems. It has only been a few years, really, since the Office of Economic Opportunity would end poverty, police chiefs would end crime, school superintendents would end reading and math problems, especially among the poor. The evaluator in many Governmental operations has been (for a number of years) the only person who was required to go out and see if these dragons came.

---

[35] See page 129 above.

By an evaluability assessment, we mean a design approach which looks at the project or process that is described by the people in charge, and looks also at the process that exists in reality. Trying to bring these two sectors together is an attempt to match up this measurable information with the questions, the goals, the objectives of the people in charge. It is true that you may find their objectives (not the people, of course) very fuzzy. You may find both the objectives and the activities very fuzzy. But by working with those people in charge and with the theory about what is supposed to work and how it is supposed to happen until the rhetorical purposes of a particular Government activity are reduced to a series of evaluable statements, you have half your problem solved. In many cases, we see evaluations where people then go to the field; and they try to assess (but there is a lot of argument in our own group about whether you should go to the field and assess at that point) whether those evaluable statements are true. If the activity in the field, on the other hand, is really quite different from the rhetoric, there are a lot of cheaper ways--than formal evaluation--of finding out how different rhetoric and activity are. A smaller, cheaper study where you try to collect that information is one of those ways. It is also a lot less visible than going out and doing a massive evaluation and finding out that the implementation is quite different, even though it may be either good or bad.

So the other half of evaluability assessment consists of recording carefully the service process or direct intervention that is actually being made and attempting to create a measurement model of the real activity of a project. This is carried out so that what is actually being done can be described in the most mundane and concrete way you can find. From this, you can assess what in reality can be measured, what those measurements would be, how they would be taken, how much they would cost and exactly where they would be obtained. By now, you anticipate my next step.

The end result of an evaluability analysis is an attempt to marry these two sets of information together and see if you can match up the potential answers that you can get with the potential questions that everybody is interested in.

We now refer to two new types of error. We not only have Type I and Type II errors;[36] we now also have Type III and Type IV errors as well.

Type III error is going out and measuring something that doesn't exist and coming back with numbers about it.

Type IV error is going out and measuring something very well, but not getting any of the things that anyone is interested in.[37] When you go to that big decision meeting in the sky or you try to distribute the information, you find that you have measured a lot of information about a real activity; but none of the things are interesting to the people who are in the discussions about what is to be done with them.

We will say if you only have two hours to design an evaluation, spend one hour on the rhetorical program and one on the actual direct

---

[36] Editor's Note: Type I error: the rejection of a true null hypothesis (that is, obtaining a statistic indicating there has been an effect, when there is no effect).

Type II error: acceptance of a false null hypothesis (that is, obtaining a statistic indicating there has been no effect, when there is one).

[37] Editor's Note: These problems are discussed at length in the Urban Institute's Working Paper 783-34, "Evaluability Assessment: Avoiding Types III and IV Errors," John W. Scanlon, Pamela Horst, Joe N. Nay, Richard E. Schmidt, and John D. Waller, January 1977.

intervention. If you have two days to design an evaluation, try spending one day on each job. If you have two months, spend one month on each job.

It is not so much that there is a fixed cost to evaluability assessment, but that there must be a fixed attitude of these continuously recurring attempts to match the answers to the questions and the questions to the answers. Because you are really trying to design a workable path for producing information out of what is going on and bringing it back to the people who are in charge of it. We put great stock, as you can tell, on bringing information back to the people who are in charge of it, even if they don't want it.

At the same time, you are really getting a lot of the basis for a technical evaluation design. We don't view this effort as a prelude to evaluation. We really view it as a use of evaluation tools in producing information, although people will make a lot of arguments about the level of belief; but I think those are philosophical arguments. There are many ways of producing things that are just beyond question (or beyond belief!). Unfortunately, a lot of those academically sure ways do not work very well in actual complex programs. There are a lot of ways of producing less convincing proof that can be applied pretty well. You are always in a trade-off between what is possible and what is desired in a real program and a real program evaluation.

The typical local criminal justice administrator needs to know more about a new approach than that outstanding people under a particular set of conditions (which are generally different from their own) were able to do it successfully. We believe that before gambling on an approach, an administrator needs to know if it has been successful in a variety of settings when operated by ordinary people. In this sense, the broad block grant program is pretty good. If you can collect a lot of these projects in a topic area and they're being operated

by ordinary people in operational agencies at the local level something may be learned, whether it's in the court or police or corrections or diversion programs. What did we send Phase I grantees out to do? The work description is available also.[38] Call Dick Barnes and get the work description. Somebody described it last night as a spiral staircase.

The NEP Phase I study tries to introduce a short intense prior step, a form of evaluation design that includes the synthesis of measurement models for the area under consideration, collection and assessment of the information that is available so you can try to see what is known, what will need to be known and what is knowable. Don't forget that last step. You may find yourself in a position of promising people answers that simply aren't knowable from the programs that exist.

By going step by step and exploring what is known, we feel that a quicker overview can be provided. Unnecessary errors can be avoided in design or evaluation, and a file can be created on a topic area as you go along. One of the toughest underlying concepts to implement in these studies grew out of evaluability assessment. A conscious attempt was made to meld together the theoretical thinking in a topic area, what actually occurs in field operations, and the methodologies of measurement and evaluation. Tom White, who is here today, says that most of the one-person problems have been solved. There have been enough bright people around long enough that most of the problems that one person can solve have been taken care of. A lot of the problems today are team problems. You don't find very many people who are awfully good in theory in a topic area and who are also good in

[38] Editor's Note: The Work Description for a Phase I Study is available from the National Institute of Law Enforcement and Criminal Justice, LEAA.

the measurement and evaluation that needs to be done later. You really need to meld those skills together.

We and the grantees--probably they more than us--have found it a very painful meld. We tried to do it with a structured work description that included issue papers in the area to try to address the theory and what people said was being done and should be done. Flow and function information from actual projects in the field was also included. First, a survey of the projects (usually by telephone) and then visits to a lot of projects to try to take down exactly what intervention was carried out and how it's connected to the criminal justice system. Then we ask study teams to synthesize a framework for description and evaluation and to assemble against this framework what knowledge is already available that has been produced in other reports and what knowledge they picked up on their field visits. In other words, they are to call out in terms of the framework and the issues what everyone wants answered, what gaps there are in the knowledge and how they might be filled. They are also asked to try to design the measures and the approaches they would use, if they had to look at a single project in this particular topic area. I will give you a list of topic areas later, but they are quite diverse. The work description had to be fairly general.

There was a lot of argument at the beginning about how much this should cost and how long it should take. Arguments ranged from $20,000 in four months to hundreds of thousands of dollars in years. We finally settled on a kind of a nominal size which varied little with the different topic areas. LEAA shot for a six- or eight-month turn-around which proved to be, I think, too optimistic; and certainly most of the grantees who are here will feel that that was too optimistic.

We kept track of it all as they went along. After running the first batch through and looking at them, we knew a lot more about the

process. Each of the 19 full reports completed has been read by all of the members of a team made up of LEAA and Urban Institute people, and each has been rated as to its Phase I-ness and probable usefulness. We have kept at it until we have gotten forced-choice paired-ratings on several criteria. As Dick said in one of our meetings, "It's a very select game. In order to come to the table and play, you have to read all 19 reports." One of the reports is 1,800 pages long. Some of them are shorter than that.

The early Phase I study leaders' comments and complaints were all gathered and combined. We took a lot through interviews and a lot through meetings that we had at different times with people doing the work. These were combined with the ratings of the study, section by section, to try to get information to rework the work description. When one of these things goes right, you are not exactly sure what has happened; and when one of these things goes wrong, you don't know quite whether you made an error in explaining it, whether the topic area is sort of impossible, or whether the grantee has fallen on his face. With a sample of 19, obviously I'm not going to say we have experimented and will determine the critical five or six factors that are in there. But I will say that we are keeping track of them, and we are trying to feed them back now into what the agency is doing so that they can do a better job on the next ones that they do.

We are using phone surveys to follow up the summaries that are distributed. I didn't bring any summaries with me, but there are small summaries that are being distributed nationally. We are doing phone surveys of local and state people to see, did they get it? Did they read it? What did they think of it and can they tell us anything they have done as a result of it or anything they are going to do?

We are using interviews to follow up actual users in the agency. There are several of the studies that have actual line users in the agency, and we are going to interview them. We already have done some interviews to follow up what they think they got out of the study. So we are trying to put all of this together and address this question of joint levels of use, the question of what is effective information to put out. There is one thing that people were saying yesterday which is very true--that the higher you go in an agency, the more people want and need one-line descriptions. When Congress improves their oversight, this problem will, of course, go away. They will be ready to take complicated textured information about textured programs. But until that happens, the higher up you go, the more you need something that is almost a press-release level of information about the study. I think it has been very hard for the grantees because they know that their information is going to be reviewed at various levels and they can almost predict at different levels who is going to be happy with it and who is going to be unhappy. Nevertheless they have gone ahead drawing up their summaries. And LEAA took a policy quite early that not only did they not want to affect (if they could help it) what their grantees put in the summary as far as conclusions were concerned, but that they didn't even want to give the appearance of affecting it.

The Urban Institute reviews each product as well as LEAA. If we think the summary doesn't match the content of the full report we send them an advisory, and we say, "Hey, we don't like this part of the summary because we don't think it matches what's in the report." There is a regular process for convening, meeting and having an argument about that. But the further up you go, you do have to reduce the amount of information; and there is more and more pressure to have a result that matches what people previously told people they

are doing and what people previously told people the results are. However, the grantee's own final summary is made available in each case.

We have some difficulties in deciding how well we are doing in terms of study quality. If you let 19 studies and you know what you'd like to get out of them, how many of them should be good? We do have informal knowledge of other people's internal reviews of sets of studies where somebody in some agency has looked at the research that they have bought. Generally, if 50 studies are examined, say, some of them can be eliminated. That is, they are not any good at all. Another batch of them may have usefulness, and another batch of them are really useful. Generally, the figures that I have from various agencies run about 35 percent, if you want to take a middle range of how many studies turned out. That is, 35 percent of all studies let are really useful. Unfortunately, not enough of these studies of buying research have been done systematically, and not enough have been done in an open way where you can use them for comparison. There is still enormous pressure on people in Government to say that every grant that they let produces something.

If you are not in Government, you can say, I am going out and I am going to let 50 grants and I expect two-thirds of them to go sour. If you are in industry, you can do that with your research; nobody expects all your research to pan out. But in Government there is still this feeling that all grants should be perfect; they should all come out. If anyone here should happen to know of any comparisons that I can use on yields of contract research, I wish you would see

me some time in the next few days because I only have one or two comparisons now that I can publicly use. Three or four that I thought I could use have been withdrawn by people who called up and said, "Gee, when I gave you that letter, I gave it to you for your own use; and I really don't want you to use it as an open comparison because nobody here will understand." We are really having trouble grappling with that issue of what kind of yield you should get out of a set of studies like this. We are going to try to treat it in the case study, so if you all have examples that you know of, that I can use for comparisons, I'd appreciate them.

I will just run through the topic areas of the first 19 Phase I studies. They were Neighborhood Team Policing, Specialized Patrol, Traditional Patrol, Crime Analysis, Pre-trial Screening, Pre-trial Release, Youth Service Bureaus, Prevention of Juvenile Delinquency, Juvenile Diversion, Alternatives to Juvenile Incarceration, Detention of Juveniles and Alternatives to Its Use, Project IDENT, Citizen Patrol, Citizen Reporting, Early Warning Robbery Reduction, Premise Security Surveys, Treatment Alternatives to Street Crime (which is a drug treatment referral program), Court Information Systems, and Halfway Houses.

Let me anticipate a question by saying that when we took this approach to the topic of Prevention of Juvenile Delinquency, nobody thought that anyone would come out with a complete framework for juvenile delinquency prevention. But it was an area of examination that was just getting on its feet. The agency had to have some tools to go in and explore it. Because this was a structured approach, they pushed some people into it to do some early exploration from which they could use the data and information that were produced in their continuing work.

I think that is about all. I am ready to open up for questions.

END