On Measuring Public Sector Performance

Mark H. Moore

Talk to the EVALUATION TASK FORCE

Of the Harvard Initiative on Children and Schooling

(Originally Delivered: December 3, 1997

(Edited for Publication: March, 2002)

I. Introduction

Thank you very much, Carol, for inviting me to give this talk. The question of how best to evaluate the performance of public sector organizations has been a pre-occupation of mine for more than two decades. By "public sector organizations" I originally meant *government* agencies. Since I have become the Director of the Hauser Center for Nonprofit Organizations, I have included nonprofit organizations in the wider conception of "public sector organizations" as well.

The reason I have been pre-occupied with evaluation is that I believe that improved performance measurement holds the key both to managing these organizations much more successfully than we now do, and to improving their standing with the public at large. It is only by reliably and accurately reporting their accomplishments to citizens and donors that government agencies and nonprofit organizations can reclaim their lost credibility. It is only by measuring their performance that government agencies and nonprofit enterprises can find ways to improve.

So far, I've been acting as though the general subject of *evaluation*, and the more specific subject of *program evaluation*, were the same as the subject of *performance measurement*. But what I'd like to talk about tonight is what I consider to be the problematic nature of the relationship between these two subjects: "program evaluation" on one hand, and "performance measurement" on the other. Obviously, these subjects are related to one another, and both could be folded into the general subject of "evaluation." Yet, I want to argue tonight that these subjects are importantly different as well, and it is important for us to begin to recognize the differences as well as the similarities.

I take up this subject with no small amount of frustration. When I began my studies at the Kennedy School in 1969 as Fred Mosteller's student, I thought that by this time in our nation's history we would surely have cracked the problem of evaluating government programs and measuring non-profit performance. I assumed with the optimism of youth that we would now be operating high performing public and non-profit organizations, guided by powerful measures of the value they were creating for society. I find it extremely frustrating that here we are, nearly

thirty years later, still wrestling with the same inadequate mechanisms for measuring performance in the public sector.

Now, I have a particular explanation of why that might be true. The difficulty, though, is that my explanation is very controversial. And since I don't know you all very well, I'm going to creep up on you with the answer rather than give it to you straight out. So, eventually, I'll try to say it, but only after I've prepared the ground a little bit.

II. Two Different Perspectives on "Evaluation:" "Program Evaluation" and "Performance Measurement"

I'd like to approach the subject of evaluation and performance measurement in the public sector from two different vantage points. The first vantage point -- a perspective I will label the *program evaluation perspective* -- is essentially rooted in the aspirations and methods of social science and statistics. The concepts and techniques central to this perspective the ones I learned at Fred Mosteller's knee: policy analysis, program evaluation, cost-effectiveness analysis, and cost-benefit evaluation. I, and I suspect many of you, were trained to think of the subject of "evaluation" as an application of statistics to answer the important practical question of whether a particular policy or program had accomplished the purposes it was designed to accomplish.

But there's a second approach to evaluation that I've learned about as my interest in *management* (as opposed to policy analysis and design) has grown. Let's call this different approach the *performance measurement perspective*. This approach starts from a very different base, I think, than the *program evaluation perspective*. This perspective is rooted in the field of management and business administration. It emphasizes the behavioral power of financial measures of "the bottom line" to help private sector organizations meet the external demand for accountability from their investors, and to guide and motivate organizations to improve their performance. This approach relies less on statistics, and more on the techniques of finance, accounting and management control. Table 1 presents a simple, stark distinction between the intellectual roots of these two different perspectives.

Figure 1

Figure 2

Table 1:

Two Different Perspectives on the Evaluation of Public Sector Performance

"The Program Evaluation Perspective"	"The Performance Measurement Perspective"
Social Science Methods	Management & Business Administration
Statistics	Operations Analysis / Control Theory
Policy Analysis & Design	Strategic Planning / Mgmt. Control
Program Evaluation	The "Balanced Scorecard"
Cost Effectiveness Analysis	Financial Statements
Cost Benefit Analysis	The "Bottom Line"

What I've been struggling to do over the last several years is to try to bring these two different perspectives -- the *program evaluation perspective* on one hand, and the *performance measurement perspective* on the other -- into some kind of coherent relationship to one another. The reason is that I am now primarily interested in the management of organizations: how to get them to perform well. What I have found, however, is that even though I believe deeply in the technical power and social utility of the concepts of program evaluation and cost-effectiveness analysis, it is almost impossible to get these techniques reliably and systematically used in either governmental or nonprofit organizations. These organizations simply do not want to pay for or do this kind of work. Even when they do commission or carry out such evaluations of their activities, they do not always use what they learn.

In fact, the last time society revved up to do this work of using program evaluations to run organizations was in the late '60s when the Kennedy School was formed. President Lyndon Johnson had mandated that organizations throughout the government adopt PPB (Program Planning and Budgeting) as a strategic planning process. That system quickly collapsed of its own weight. Then it collapsed again when Jimmy Carter adopted ZBB (Zero-Based Budgeting) from Texas instruments. Eventually, the federal government ended up with something that looked like it might be interesting. That system was called MBO (Management by Objectives). It designed measures to focus attention on specific objectives the government sought to achieve. I think that system has now been re-created and given legislative sanction through GPRA, the Government Performance Review Act, which requires all government agencies to define the objectives to which they will be held accountable. And it may be that in the relative success of MBO and GPRA lies something that will turn out to be the right synthesis between the two distinct perspectives I have described.

But as a once devoted student of Hegel, I learned the importance of dialectical thinking. So, instead of leaping to the synthesis, I'm going to start with a dialectic tension between the *program evaluation perspective* on one hand, and the *performance measurement perspective* on the other.

III. A Close Look at the "Program Evaluation Perspective"

The program evaluation perspective is rooted in several fundamental commitments. One is that evaluation, if it is done well, should concentrate on measuring ultimate *outcomes* not processes or activities. By *outcomes*, we usually mean the ultimate intended or desirable results of a policy or program; not those effects that happen closer to the boundary of an organization, or some way along a causal chain that leads to the desired results. The program evaluation perspective wants to know whether we did or did not get to the final result we sought.

A. The Value Chain: Inputs, Processes, Outputs, and Outcomes

Figure 1 is a simple diagram that describes a flow of production and causation. In the business world, this would be described as a "value chain." In the public sector, this would often be described as the "logic model." The people who worry about these technical relationships in the private sector would be called production engineers. The people who worried about these issues in the public sector would be called program designers. In both cases, the general idea is that in order to produce results, we have to have some particular ideas about the process by which fungible inputs can be combined in particular ways to produce particular desired results.

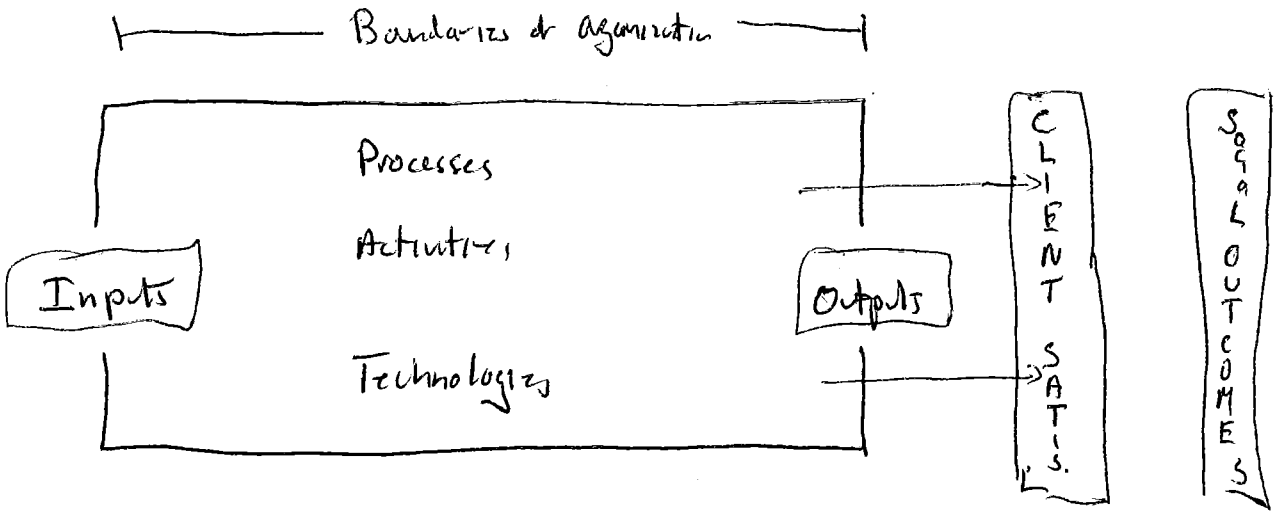
I assume that all of you have own version of some chart or some figure like this. You may put different words on different parts of this flow of production and causation. But while we might argue about which words to use, I think we can all agree that there's some kind of production process that starts with "inputs" and then uses those "inputs" in certain "production processes" or "activities," to produce "outputs," and the "outputs," in turn, are expected to produce "outcomes."

B. Organizational Outputs versus Desired Outcomes

The principal reason I present this Figure is to make it clear the special way in which I'm going to use the words "output" and "outcome." "Outputs" are activities or events that occur *at the boundary of the organization.* They are the things a manager can see happening when an employee of the organization is interacting with clients or with other features of the world that the organization is trying to shape. "Outcomes," on the other hand, are things that happen further

Figure 1

The Value Chain in Public Sector Operations



down a causal chain -- at a place remote in space and time from activities carried out and sustained by the organization. What is important about *outcomes* is that they are the results that constitute the ultimate value that the organization is trying to achieve.

A concrete example of the distinction might be helpful. In examining a methadone maintenance program, we could say that the *output* of the program would be administering methadone to eligible heroin addicts under a regime that included urine surveillance, job training, psychological counseling, etc. The *outcome* of such an effort would be the changes in the health and employment status of the addict, as well as changes in levels of their criminal activity.

I belabor this point because some analysts (particularly economists) do not make this distinction. They use the words "output" and "outcome" as essentially equivalent. They treat both as meaning the "value" created by an organization's activities. Both are set up in contrast to both the resources that are used to create this value (the "inputs"), and the processes or technologies that is relied upon to convert resources into outputs (the "production function"). Because the valued results that an organization produces always lie "downstream" in a production process, the outputs of the organization – the value added by the organization's activities -- can be easily distinguished from both inputs and production processes.

For my purposes, however, I need to make a distinction between "outputs" on one hand and "outcomes" on the other. To repeat, outputs are things that happen right at the organization's boundary. "Outcomes", in contrast, are things that: 1) happen remote in space and time from the organization and 2) contain the essential rationale or the value or the justification for the particular enterprise.

This distinction is particularly important in the "program evaluation perspective" for the simple reason that this perspective demands that we focus our attention on "outcomes" rather than "outputs" or "processes." The reason for demanding a focus on "outcomes," of course, is that "outcomes" are, by definition, the results that provide the *raison d'etre* for the enterprise. We can't be sure, either for policy-making or managerial purposes, that anything of worth has been achieved unless we can point to the "outcome" that was achieved. To know confidently whether

something of value was created, we have to be able to locate and measure not just the process and the output, but also the outcome.

Now, I'm going to pause here for a minute to hint at the potential for a very long and interesting digression. Note that what is important about the idea of an "outcome" is that it is a result (or less restrictively, simply an occurrence) that we think is *socially valuable.* I emphasize the idea of socially valuable to underscore the fact that there's a normative idea there. That idea is that *"we"* say that *we* value some particular set of events or effects or results occurring in the world. To the extent that *we* thought these effects had occurred as a result of a policy or an organization's efforts, we could conclude that the policy or organization had created value. In this sense, "outcomes" are to public sector organizations what revenues earned by the sale of goods and services are to private sector organization: namely, measures of the gross (but not the net) value created by organizations.

C. Outcomes as "Public Value"

By thinking about outcomes as the value created by public sector organizations, one suddenly realizes how much is at stake in defining the outcomes to be measured in program evaluations. These become the dimensions of value to be considered. The definition of the outcomes is the functional equivalent of the development of an accounting framework that allows us to recognize (in a strict accounting sense) value creation.

One also realizes how important the development of this accounting framework would be for the successful management of public sector enterprises. To define and recognize value is a pretty fundamental idea. How could one run an organization if one didn't know what was valuable to produce? So, it must be that there's some concept about how one would define and recognize value in order to accomplish it.

D. General Methods for Defining "Public Value"

Yet, when one considers (as I have) the history of methods we have relied on to recognize the public value created by public sector organizations (and here I am talking primarily

about government organizations), one finds that we have relied on only four methods to define what constitutes public value. Table 2 arrays these ideas in chronological order.

1. Public Value Defined by Policy Mandates

The first method says that "public value" is what elected representatives of the people acting through legislation or executive orders say it is. When the elected representatives assert that some goal is important to achieve, we have to treat that as a normative declaration by the broader society about the purpose that they intend through the governmental organization.

Note that while one might scoff at this idea on grounds that one could hardly expect the messy and corruptible processes of politics to be capable of defining something of public value, it is precisely this idea that often provides the starting point for program evaluations. After all, program evaluators have to start with *some* idea of which of many possible things to measure in the world as important effects or results of a program that should draw their attention. They do not wish to import their own values into the analysis. So, they often begin with an idea of what the legislatures or policy-makers seem to have intended as results from the program. This gives them the dependent variables they need to measure in their program evaluation work. Insofar as they are treating the effects that legislators intended as the desirable results of the programs, they are implicitly acknowledging that it is the policy-makers' aspirations that define the value of the program.

Anyway, the first idea about public value is that public value is defined by political and policy mandates. Of course, even though we remain an enthusiastically democratic country, our confidence in our elected representatives wavers at times. Sometimes, we lose our confidence in the capacity of a political process to make an intelligent choice about what constitutes public value. At various times in our history, then, we've tried to construct alternative methods for defining value that could stand against political decision-making, or be added to political decision making.

2. Public Value Defined by Substantive, Professional Expertise

The first method we developed during the Progressive Era was to try to substitute substantive professional expertise for political judgment. In this view, if you wanted to know how to construct a sewer system capable of protecting an urban population from disease, the last person to ask was a politician. The right person to ask was a sanitation engineer. Similarly, if you wanted to know what kind of an army we needed and how much of it, the right person to ask was a general, not a politician. If you wanted to know what kind of education would be good, you asked teachers and principals; you didn't ask politicians. And so forth.

Confidence in the substantive expertise of professionals in judging value is still both relied upon by politicians, and invoked by substantive professionals. But gradually, that star, too, has been tarnished. We lost our confidence in substantive expertise because we learned that substantive expertise often became self-serving. We discovered that substantive experts were inclined to "gold-plate" their proposed solutions. The frequent result was that society ended up paying more than they needed to produce a particular result. On occasion, the experts found ways to benefit themselves using their expertise as a screen for sloth or greed.

3. Public Value Revealed Through Analytic Techniques

To deal with the potential corruption of both politicians and substantive experts, then, we launched a major effort to create and deploy analytic techniques that could tell us objectively whether government and nonprofit organizations were creating "public value." The hope was that we could find some measures that were as objective and compelling as the "bottom line" in the private sector. This occurred at about the time I came of age. In fact, these were the techniques that I came to the Kennedy School to learn: namely, the techniques of program evaluation, cost-effectiveness analysis, and cost-benefit analysis. We thought, at the time, that these would be very powerful instruments. We thought that these techniques would do for people in the public sector what the ordinary tools of financial analysis did for people in the private sector. The tools could be used to reveal whether an organization as a whole was creating value, and what particular parts of the organization were performing well and which badly.

The promise of these techniques was great. Yet, I think it's fair to say that we've all been a little bit disappointed in the success that these techniques have enjoyed. We'll come to some of the reasons why. But before we talk about why these techniques have been less successful than hoped, I want to quickly review the last idea about "public value."

4. Public Value as Customer Satisfaction

Our most recent idea about how to judge the value of public sector enterprises has been that we ought to measure value the way that private sector organizations do: by measuring the satisfaction of the "customers" of government. Now, I often go over to the Business School and talk to my friends there about managing organizations. They often say to me, "You know, Mark, government is just a very large service enterprise. What you need to do is understand what we know about how to run service enterprises. If you understand that, you'll be able to make the government perform better."

I confess that as they're talking about service delivery organizations, I'm thinking about the organizations with which I work: police, prisons, tax collectors, safety inspectors, and environmental regulators. And I'm thinking to myself "Should I think of the clients of those organizations as 'customers'? It doesn't seem to me that these people are being served by the organization -- more like that they are being serviced! They aren't paying for the privilege of being investigated or inspected. And it isn't obvious to me that the point of these organizations is to make these people happy, though it might be desirable to do that. Those who are arrested and imprisoned by the police to control crime, or forced to pay taxes to support public purposes, or regulated to ensure environmental quality are people who get *obligations*, not services."

In fact, the more you look at it, the more you realize that most services in government come with obligations. And the reason that's true is because part of what we're trying to do, often, with our clients in governmental organizations is not only do something that's good for them but by doing so accomplish a purpose that was originally set by the political aspirations. In that important sense, then, our clients and customers often become the means to achieving some legislatively established end, rather than ends in themselves.

E. Outcomes as Socially Defined Purposes and Clients as Means to Socially Defined Ends

The reason I go through all that is to help us notice that the concept of an "outcome," which is so fundamental to both program evaluation and cost effectiveness analysis, is precisely a concept that treats the client as a means to an end. When we evaluate the success of a methadone maintenance program, for example, we're not simply asking the question whether the addict/client liked the program. We're asking the question whether the person came regularly and improved in terms of reducing their crime and increasing their employability and increasing their parenting skills as a consequence of that. The reason we're doing that is that the valuing of that result is being done by the politicians representing citizens, not only by the clients.

F. Why Program Evaluation is Superior to Cost-Benefit Analysis

Now, the idea that outcomes are defined collectively as socially desired results, and that we use those outcomes to evaluate governmental programs, is, I think a very important idea. And while it is consistent with the (largely statistical) training of those who do program evaluation, this idea is inconsistent with the (largely economics) training of those who are trained to do cost-benefit analysis.

I have to tell you that it is wonderful to be in room filled mostly by statisticians rather than economists. At the Kennedy School, it is usually the other way around -- with more economists than statisticians.

There, what I hear, is that the only correct way to judge the value of governmental programs is through cost-benefit analyses. Moreover, I am told that cost-benefit analysis does not mean something sensible like: "Well, let's imagine all the attributes of performance that you could consider as important effects. And we'll call the good (positively valued) things 'benefits' and the bad (negatively valued) things 'costs.' And then we'll figure out some way to decide whether the 'benefits' outweigh the 'costs.'"

They mean something much different than that. They often mean that the right way to evaluate a government program is not against some kind of "social maximand" established by the legislature. They think that the right way to evaluate a program is allow each person in the society

to record their experience -- positively or negatively -- to the effects of a policy, and then sum the positive and negative experiences that each individual has over all individuals to determine whether the effects of the policy were positive or negative. In short, instead of having us all talk about what we would like to accomplish together, and on the basis of that collective process write down our collective ambitions, and use those ambitions as a social maximand, they think the right basis is for each of us to experience the policy, and say whether it was good or bad for each of us alone.

It turns out we should blame Jeremy Bentham for the economists' idea about how to value social programs. I was quite astonished to discover that the opening paragraph of "Utilitarianism" sets out Bentham's purpose in constructing his theory of utilitarianism. That purpose was to give guidance to legislatures that had to make difficult public choices about what was to be done on behalf of the public. He was our first policy analyst. He was worried about the capacity of the ordinary political process to make good decisions. He wanted to make sure that there was something that could discipline the ordinary political process. So, he invented utilitarianism as a guide to the legislature as to how to make decisions that would be in the public interest. And the public interest, of course, was that set of things that maximized each individual's welfare; not those things that a collective political process decided were important to achieve regardless of what individuals who were affected by the policy thought.

Now, unlike economists, I believe in the idea of *e pluribus unum*; that is, I believe in the capacity of collective political processes to transform individuals into a "we" that can value things as a collective. Therefore I also believe in the possibility of constructing social maximands as weighty normative claims on government agencies. If I'm right to believe in all this, I come to a conclusion that should hearten statisticians trained in techniques of program evaluation. The conclusion is that program evaluation and cost effectiveness analysis are not the poor relatives of cost benefit analysis that serious analysts have to resort to because cost-benefit analysis turns out to be practically impossible (though theoretically desirable). Quite the contrary! Program evaluation and cost-effectiveness analysis are conceptually the *right* way to approach the subject

of measuring the value or public programs, and are to be preferred over cost-benefit analyses done the way the economists recommend. The reason is that both program evaluation and cost effectiveness analysis are based on the idea that value lies in the collectively mandate purposes of the program -- in the social maximand that defines legislative intent for program evaluation, and the dimensions of "effectiveness" in cost-effectiveness analysis. Since these (rather than the simple summation of individual preferences) are the proper guides to value creation in the public sector, these techniques are to be preferred theoretically as well as practically. In doing program evaluations and cost-effectiveness analysis, the satisfaction of clients and customers are less important than one might think.

At any rate, that's a long digression that we probably didn't need to go through. But it might turn out to be important when we come back to these subjects later.

To summarize, then, one fundamental idea in the "program evaluation perspective" is that we should measure "outcomes;" not client satisfaction, not outputs, and not processes. The reason we should measure outcomes is that these are defined by the collective as the valued effects to be produced through the use of collectively owned assets.

G. The Importance of Causal Attribution

The second fundamental idea in the "program evaluation perspective" is that a crucially important task in evaluating government programs is to be sure that we can reliably attribute valued (or dis-valued) effects to the operations of a program, policy or organization. It is the emphasis on reliable causal attribution that makes knowledge of statistics so central to the *program evaluation perspective*. The reason is that we must turn to the methods of statistical inference to establish a reliable causal attribution of the desired effect to the planned activity.

Now, one could reasonably ask, I suppose, why it is that we're so interested in an accurate attribution. I'll argue in a minute that business turns out to be relatively uninterested in the question of whether it gets an accurate attribution of results. We know, being statisticians, that it is incredibly expensive and difficult to make a reliable attribution of effect to cause. We have to go through many contortions to ensure that mere experience can be transformed into

experiments that allow us to make reliable causal attributions. So, the question is: why do we think that causal attribution is so important?

The most powerful answer (it seems to me) is that these methods will produce *scientifically based knowledge* about whether a particular policy or program *works*. If that is an important goal of evaluation -- to produce robust, general knowledge of what works -- then that goal can only be achieved by getting the attribution right. Without reliable causal attribution, one cannot develop the general and robust knowledge that is at least part of the goal of program evaluation. We want to know not only that what we are doing seems to work now, but also that it will work in other times and places in the future. That generality and confidence, after all, is what social science is all about. And when we apply the scientific spirit of social science to the practical goal of measuring the results of public sector enterprises, we ought to be equally interested in producing general, confident knowledge as well as a kind of contingent operational knowledge.

A second possible reason to focus so intensively on getting the attribution right may be that we have a Calvinist desire to make sure that no politician walks off with any extra credit. We're so cynical about politicians trying to appropriate everything for their own credit that it seems right to be stern with them about whether they can reasonably claim credit for a valuable social result.

Now, one of the fields in which I work is crime control. In that field, you can sense the fury that the criminologists feel about Mayor Giuliani and Commissioner Bratton swaggering around, acting as though they have succeeded in controlling crime. It seems that they are trying to claim an entirely unearned credit. Without doing the work of demonstrating to the satisfaction of statisticians that their initiatives succeed in controlling crime, they have no right to the credit they are taking.

Whatever the reasons, I hope I am reflecting your views when I say that accurate attribution of results to causes is pretty central to the perspective of program evaluation. It is toward that goal that much of our technical expertise is deployed. And it is that standard against which we judge the quality of many program evaluations carried out in the world.

H. Limitations of the Program Evaluation Perspective

Now, having set out the two key ideas in the *program evaluation perspective*, I want to argue that there are some fundamental limitations to this perspective. One limitation is particularly important if you are interested in managing organizations: namely, that the focus of program evaluations is generally speaking not on an *organization*, or even a sub-unit of an organization. The focus of evaluation is, instead, is instead a *policy* or a *program*. That may seem to be unimportant, but I think it turns out to be critically important.

1. The Difference Between Programs and Organizations

Now, you could say "Oh, that's not true, Mark, don't worry about it. There isn't much difference between evaluating a policy and a program on one hand and evaluating an organization on the other. After all, all organizations are either (a) the implementing mechanism for a policy or (b) a bundle of individual programs. Since organizations are nothing more than combinations of policies and programs, if we can evaluate policies and programs, we know how to evaluate organizations, as well." Unfortunately, I don't think this claim is true.

I think organizations are a little bit different than the implementing mechanism for a policy. I also think they're different than a bundle of programs or product lines. What organizations are are pieces of institutional and human capital. At any given moment, those bits of human capital are committed to producing particular things. And it is in this sense that one could say that organizations are either bundles of policies or the institutional means for implementing particular policies.

But the point is that organizations are not just instruments for achieving purposes that were set once and for all by others. Organizations are more active and dynamic than that. They see opportunities in their environments, and do or do not exploit them. They learn how to do their work better, and do or do not invest in the deployment of those new methods. They think, and adapt, and innovate, and invest as well as produce. And we have an interest in how creative and dynamic they are. So, I'm not sure that we get an accurate picture of the value of an organization and its performance if we deploy technologies that are primarily good for evaluating policies and

programs and not for recognizing the value that organizations are creating. It may be that it is important to look at organizations and evaluate them not only in terms of current performance, but also in terms of the way they are adapting and learning for the future.

2. Limited Utility for Organizational Feedback

Let me just go on to the next limitation of the "program evaluation perspective." It is this: program evaluations focused on outcomes and designed to ensure the reliable attribution of results to the program are incredibly *expensive*, and incredibly *slow*. These features are necessary and inherent in the idea of program evaluation.

Two things make program evaluation very *expensive*: first, the outcomes happen remote in place and time from the boundary of the organization; second, that is administratively expensive to set things up in ways that allow accurate attribution. If we decide to measure outcomes that are remote in space in time from the boundary of the organization, we have to pay extra for measurement. If we set up the activity in ways that are designed to yield reliable attributions of effects to particular causes, we have to manage things very carefully, and give up some flexibility in adapting operations in response to the results we seem to be getting but have not yet registered in a scientific way. In short, we have to wheel out the expensive apparatus of measurement and social science for each program evaluation we wish to conduct.

The reason that program evaluations are often *slow* is not simply because researchers are slow (though we have to admit to some of that); it is more fundamentally a consequence of the fact that the ultimate effects of public programs often take some time to register their ultimate effects; for example, we may not know the benefits of enriched pre-school experiences until we see how the kids who received such assistance behaved as adolescents. Alternatively, it might well be that the effects of a program are only worth having if they are sustained over time; that is, if the drug addict stays off drugs, or the former welfare client manages to stay on the job over several years. If it takes a long time for an effect to occur, or if the effect becomes more valuable as it is sustained over time, then no matter how speedy the researcher, we must still wait for the important results to come in.

If program evaluations are both expensive and slow, that usually means, from the perspective of organizational management, that, at any given movement, organizational managers will have only spotty coverage of the value of different parts of their organizations. They won't have comprehensive evaluations across all the activities of their organizations because they will not be willing to pay for evaluations across the board. They won't have them quickly, because they have to wait for the results. The consequence, then, is that managers cannot know in real time when decisions have to be taken as to whether the effects are positive or negative.

IV. Some Alternative Uses of Performance Measurement

What might have occurred to you by this time is that there might be some different reasons to measure organizational performance than to generate general, robust knowledge of what particular methods work to create particular results. While it is always good to have and to generate this kind of knowledge, performance measurement systems are used in organizations for many other, less scientific purposes.

A. Accountability and Motivation

One alternative use of performance measurement is for accountability and motivation; that is, as part of a behavioral system that uses measurement to decide whether the organization as a whole, or its constituent parts are doing well or badly. Measurement gains motivational power in this context when it is tied to organizational responsibility, and when rewards and punishments depend on the measured performance. That is one very powerful and important use of performance measurement in an organizational context.

B. Resource Allocation

A second use of performance measurement is to guide resource allocation. The idea is that performance measurement can show us what activities of an organization are relatively high value and which relatively low. Knowing this, we can increase the overall value of the organization's output by shifting resources from low value uses to higher value uses. Note that this effect is produced without necessarily getting any motivational power out of the measurement

system. We assume technologies are fixed -- not influenced by performance incentives -- and that the value comes from allocating fixed resources more efficiently and effectively across the fixed technologies which represent alternative uses of the organization's resources.

C. Knowledge Development and Learning

The third possible use of performance measurement is to use it to test our theories and learn about which of our methods are working well. This is use of measurement for continuous improvement. This may sound like the development of scientific knowledge, but what I have in mind here is a kind of measurement system that simply tells us whether things seem to be improving in one or more valued dimension of performance. This is less than knowing for sure that a particular thing the organization did can be relied upon to produce the same valuable results over and over again.

D. Differences Between Public and Private Emphasis on Purposes of Performance Measurement

Now, I've arrayed these different uses of performance measurement in approximately the order that business thinks about the right use of measurement. They think accountability and motivation is much more important than resource allocation; and both of those much more important than the learning or testing or methods. My hunch is that those of us who give advice to the public sector about how to use performance measurement tend to reverse that order. We think that learning and testing of methods is terribly important, resource allocation is the next most important, and the least important thing (because it has to do with the mundane job of running organizations) is associated with accountability and motivation. I'm not sure that that's true but I just want to hold that out as a possibility.

V. A Different Perspective: The "Performance Measurement" Perspective

So, now let me turn briefly to the *performance measurement perspective*. This, as you recall, is the idea about performance measurement that comes from business. It is closely tied to financial analysis and management control (rather than marketing, product design, or production engineering). The fundamental idea is to use measurement for external and internal

accountability and motivation. That's what's important about it -- much more so than either resource allocation or the testing of production methods.

A. The Utility of Financial Measures -- Particularly Revenues from Sales

The most fundamental idea within the "performance measurement perspective" is that financial measures give powerful information about the value that's being created by the organization or the success of the business plan that has been adopted. The idea here is that there's just an enormous amount of information contained in a private sector organization's financial measures.

I admit that I've studied enough about business management to have become truly envious of the fact that they have financial measures; more specifically, that they have information about the revenues that earned from selling units of products and services to willing customers. For purposes of both evaluation and running organizations, it is just spectacularly useful for organizations to have financial information in general, and the revenues earned by the sale of products and services in particular. Indeed, I've made a list of just how wonderful it is to have information in one's organization about the amount of money the organization has earned from the sale of products and services to willing customers. Here it is.

First wonderful thing: the revenues earned from the sale of products and services are, in fact, a direct and fairly objective measure of value. (At least value at the individual level, if not at the social aggregate level.) We know that our products and services have value because individuals voluntarily plunk their money down to obtain them. That simple fact tends to end lots of discussion about whether something is valuable or not. When teaching government executives, I sometimes like to pander to them by saying something like the following: "Where in the world do those business guys get off spending important time, resources, energy on the production of hula hoops and lemon scented furniture polish? I mean, give me a break! With all the problems in the world what the heck are they doing producing lemon scented furniture polish and hula hoops?" The difficulty, of course, is that the business executives have a pretty convincing answer to that question. They say, "We produce this stuff because people like it. We know they like it because

they buy it. And, since we know that the amount they paid more than covered the costs of producing it, we know that value was created." To the extent that the collective thinks that satisfying individual desires for things is an important social goal to achieve, then one could say that social value has been created as individual satisfaction (because individual satisfaction is part at least of what society as a whole wants). So, the first thing that is valuable about a revenue earned through the sale of products and services is that it is a direct measure of value.

Second good thing about revenues: revenues perform the magical feat of comparing apples to oranges. We can start by saying you just can't compare apples to oranges. They are so different. But, when people pay thirty cents for an orange and twenty cents for an apple, we know that if we produce three oranges it is worth more than four apples!

Third spectacular feature: we can compare value produced directly to the cost of producing it without having to change metrics. The fact that individuals are willing to pay thirty cents for an orange tells us how much they value that orange. If we know (from our cost accounting activities) that it costs us only five cents (and some help from mother nature) to produce that orange, then we can conclude that growing the orange and getting it to market is creating net value for the society. The revenue gives us the gross value of what is being produced. The costs tell us how much value was used in the production of the good. The difference between the revenues and the costs gives us direct information about whether net value was produced.

Fourth really good feature: revenues earned are measured quite inexpensively, and quite quickly, right at the boundary of the organization. The organization gets the information about the value of its product at the point of sale. It doesn't have to wait to see if valuable effects occur. It doesn't have to go out and ask people at sometime in the future whether they have changed. All the information about value is right there at the moment the money passes over the counter.

Fifth excellent thing: revenues are reliably and accurately measured because we have centuries of tradition in developing systems to prevent theft. Financial management systems weren't really created to measure value creation. They were invented to prevent theft. Because

we've had these systems for centuries, however, we've developed the cultural commitments, the institutional infrastructure, and the specific technical systems to measure revenues accurately.

I go through this list because I am trying to emphasize just how powerful financial measures are as measures of value creation in the private sector. To make this claim even more vivid, just consider what would happen if we put the head of GM, or a GM plant manager in the same position as public sector managers: namely, that they could have all the *cost* information they wanted, but that they could have no information about the *revenues* they had earned by selling the cars they made. It is clear, I think, that these managers would have lost a great deal in terms of their ability to direct and control their organizations. They would now know the value of what they were producing, and could not use that information either to reassure stockholders, allocate resources, or motivate employees.

To regain some ground, they might well turn to some of the (much less satisfactory) measurement techniques that are common to the public sector. Just as we asked the sanitation engineers what kind of sewer system we needed, and generals what kind of army we had to have, we might ask our automotive engineers about the features of good cars, and whether our cars did or did not have those features. Just as we went out into the field to measure the impact that methadone maintenance programs had on the behavior and condition of heroin addicts enrolled in the program, we might go out and try to measure the performance of the cars on the road with respect to qualities such as transportation mobility, comfort, fuel economy, etc. Just as we might ask those who interacted with government for information about their satisfaction with the quality of the service they received, we would take surveys of those who bought the cars, and ask whether they liked them or not. This would all be valuable information, but we would still be missing the most powerful information of all: namely how much customers valued our cars and whether the valuation was more or less than the costs.

B. Performance Measures as the Public Sector's Functional Equivalent to Revenues

So, given the power of the information contained in revenue information in the private sector, I keep thinking to myself that something that looks like revenue information is what we need in the public sector.

This insight, I think, turns out to open an important avenue of attack in thinking about performance measurement in the public sector, and helps explain why MBO and GPRA are potentially more useful that PBB or ZBB. The basic idea is simple: that it might be important to concentrate a lot more of our fire-power -- our intellectual energy, our political energy, our organizational energy -- on *trying to develop some simple, plausibly important measures of processes and output rather than outcomes*. We need to do enough careful program evaluations to check the adequacy of the theories we are using to guide public organizations toward the achievement of valuable results. But we don't have to do evaluations everywhere all the time. We can strategically target our expensive program evaluations. Everywhere else we can and should be spending a lot more time and effort measuring processes and outputs. I want to try to defend this position in favor of output measures through two different arguments.

The first argument is that measuring outputs and production processes generates a great deal of energy inside organizations, and that energy has the potential for increasing the organization's performance if the measures are even roughly in the direction of the organization's ultimate goals. Measuring processes and outputs (as distinct from outcomes) have these effects simply because output measures happen in "real time," are closely linked to things that managers can control, and can be tied to organizational rewards and punishments.

Let me give you an example from the private sector. McDonalds and Burger King could, in principle, manage their individual franchisees on the basis of their equivalent of outcome measures: namely, measures of financial performance such as profitability. All they'd have to do is take the cash register receipts, and subtract the cost of materials and payroll. The simple manipulation of financial documents could provide information on profitability for every restaurant.

If measuring outcomes was such a good thing to do, you would imagine that McDonalds and Burger King would run their restaurants through financial measures.

In fact, they do not operate in this way. Instead, they hold their mangers accountable for twelve specific characteristics of the experience of being in the restaurant: how long was the line, was the food fresh when you got it, was there place to sit down, was the bathroom clean, and did the server smile at you when they took the money. These are all very concrete, specific measures of *output*, not about *outcomes*.

One of the reasons they measure these particular things is that it is these things that they think of as the *product* of their restaurants. They don't think of their product as a hamburger. It is, instead, the *experience* of being in the restaurant. The relevant dimensions of that product are measured through the particular system they have created. They believe (and they get a chance to test this belief relatively quickly and easily) that if they can produce that "product" (i.e. the particular experience of eating in their restaurant), people will buy it and it won't cost very much, so they'll make money.

Of course, what makes their situation different than a public managers' is that they get a chance to test that theory of value creation through the ordinary operations of the market. But the point I want to emphasize is that the private sector managers get a lot of performance out of their organization by focusing on concrete measures of output guided by a theory of value creation. In fact, they think this is so important that they spend a great deal of time and money taking the measures of restaurant performance. They measure their stores on these 12 dimensions of performance by sending a person in to act as a customer about once a week, 52 times a year. That, I think, is a much higher level of investment in measurement than we make in the public sector.

Indeed, when you report this to public sector managers, their first reaction is horror. They think they couldn't possibly put up with that level of surveillance and control. Then they begin thinking about it and realize that if top management were making measurements that often, then top management would be able to observe relatively small changes in performance. They then

think that they might actually be able to produce some relatively small changes in performance, if not big changes, if not ensure profitability. They also note that the evaluation of their restaurant is not, as it is in many public sector organizations, like lightening striking once a year. It is more like a continuous voltage running through the organization that keeps them on their toes, rather than toasts them unexpectedly.

That's what I mean by the power of process and output measurement to focus managerial attention and drive organizational behavior. The more frequent the feedback comes, the more powerful the comparisons that can be made, the more connected it feels to the culture and strategy of the organization, the more powerful these measurement systems become. If that behavioral power is available to a manager through performance measurement, it tends to push one back from outcomes toward output and process measurement.

C. Linking Performance Measures to Outcomes Through Logic Models

The worry, of course, is that by shifting from *outcome* measurement to *process* and *output* measures, we cannot be sure that we are producing much of value. It may be that we are driving the organization powerfully in the wrong direction.

That is certainly a worry. But the way to deal with it, I think, is not to try to measure outcomes all the time. It is, instead, to formulate an explicit, detailed theory about the link between outputs and outcomes, and test that theory intermittently to determine if it is correct.

Taking on these responsibilities for explicating the basic logic model that links organizational outputs to desired results is a valuable discipline even if one never gets around to testing the assumptions. The reason is that there is much that can be learned by making explicit what is implicit, and then testing that explicit theory against common sense as well as carefully constructed experimental results. Moreover, once one has laid out the logic chain, one can often find easier and quicker tests of whether the theory is correct than waiting for the outcomes to materialize. So there is great utility in accepting the idea that one must, perforce, operate on the basis of a *theory* of value creation rather than demonstrated success; and that it is desirable to be as clear as one can be about what that theory is, both for purposes of lending plausibility before

the experience comes in, and for efficiency in testing the theory. I suspect many programs would be found to be inadequate simply because they failed to meet the "giggle test" once their logic was explicated, and that the discipline of being clear about the assumptions of a program would be much cheaper, faster, and more powerful than expensive program evaluations.

Similarly, once one knows that one is operating on a theory, it becomes possible to strategically target limited evaluation resources. One can simply look across the theories one is relying on, and figure out which ones are commanding most of the organization's resources. An example from policing is useful here. Once one understands that about 70% of an organization's resources are devoted to two simple processes -- patrol and rapid response to calls for service -- it becomes very important to learn whether those processes are working to reduce crime and fear. One doesn't have to test this question over and over again in every police organization. One can do it in a few, and let the results stand for other organizations as well. That is far more important than testing the much smaller effects of smaller programs such as DARE or reliance on arrest in instances of domestic assault. Yet, in many ways it seems unnatural to do an evaluation of random patrol because we think of this as a huge organizational process, not a distinct program or policy.

D. Processes and Outputs Are Valuable In Themselves

The second argument for focusing on processes and outputs as well as outcomes is that there are some characteristics of transactions at the boundary of the organization and of the production processes themselves that have value in themselves, not just as a means to the end of achieving results. These characteristics of transactions and processes could be valuable either to the clients who show up at the boundary of the organization. Or, they could be valuable to those who are authorizing the organization's activities: namely, citizens and their elected representatives in the legislature, the executive branch, and even the courts, who are saying they want and value certain characteristics of outputs and certain characteristics of the production processes as well as particular outcomes or results. For example, again using the case of the police it might be important to both overseers and clients that the police minimize their use of

force in encounters with citizens. Or, it might be important to overseers of the police, that the police allocate their resources across precincts according to need rather than population or tax base.

In my view, these arguments together amount to an argument for much more determined efforts to measure processes and outputs, as at least a necessary if not sufficient part of performance measurement in the public sector. That's the basic argument.

E. Examples of Performance Measurement

To give the argument a bit more illustration and plausibility, let me give you two examples from my own experience. One example comes from the time that I was working for the Drug Enforcement Administration in the mid-1970s. They had, at the time, an interesting performance measurement system that they were using to run the organization called the GDEP program. Prior to the creation of the GDEP system, they had operated according to a concept they called the "system model". The "system model" required them to develop (through intelligence) a comprehensive picture of the illegal distribution systems, including a picture of who were the most important dealers. With this intelligence picture, they then assigned organizational priority to attacking particular parts of that system. Whenever they arrested someone who was on this target board, they gave themselves a lot of credit.

While the "system model" seemed a logical way of maximizing their impact on the illicit drug trade, it turned out to be hopeless organizationally. One reason is that the distribution systems were changing fast enough that their "system model" was often outmoded. Even worse, it was often true that they didn't have any way of going after the people whom they considered high priority targets. They kept grinding away at surveillance, but they simply couldn't develop the informants that would allow them to make a successful case against their key targets. In the meantime, because they were focused on the particular targets specified by the GDEP model, they kept turning down opportunities to arrest other people who were dealing drugs but were not on their board. In short, the system model focused their efforts rigidly on targets whose value they couldn't really be sure of.

They decided, eventually, that they needed to abandon the "system model," with its emphasis on particular individuals to be arrested, even though it seemed like a logical way to be focused and proactive. They shifted to a system that encouraged everyone to make arrests, but then graded the quality of arrests by the importance of the person who was arrested. Class 1 violators were the captains of the illegal drug trade. Class 4 violators were the mules and the street dealers. They developed some objective criteria for deciding each class of violators. A separate unit of the organization graded the arrests that were being made. They not only graded the level of dealer; they also graded the kind of drug. Heroin and cocaine dealers were considered much more valuable to arrest than marijuana dealers.

So, they invested in this very simple system that allowed the organization to act much more opportunistically and flexibly, but still kept it focused on value by grading the quality of the product they were producing. The system allowed them to have and operate on something that was more general than the arrest of people we had identified (on ambiguous, time sensitive intelligence data) as very important, but more specific than a homogenous product called "arrests." The system turned out to be quite powerful in driving the performance of the organization.

What turned out to be interesting, though, was that although this system was powerful in creating a sense of accountability and driving the performance of the organization, it was much less useful in guiding resource allocation decisions. In principle, the system could give you information about how to allocate agents across regional offices in the United States. If we were making many more high quality arrests per agent in Miami than in Kansas City, it would make sense to transfer agents from Kansas City to Miami. In order to make the comparisons, however, one had to decide quantitatively how much a class-one violator arrest was worth relative to a class-four violator.

Now, I had a theory that would link that output to the outcome. And in my theory the class-one violator arrest was worth like 200 or 300 times the impact on the illicit distribution system than the impact of a class-four arrest. If the organization bought that theory, and built that

into the GDEP system, when they summed the scores over all arrests for the various regions, the system would naturally produce a high degree of variability not only across regions, but also within a region over time. One big arrest could make one region look very good, while another region that had been producing a high volume of less serious cases would look pretty bad.

That was the whole point of the theory, of course. The theory was based on the idea that it was much, much more valuable to arrest a class-one violator -- a "king pin" -- than a class-four violator -- an easily replaced "mule." So, the point of the system was to reward regions for the value they created, not the effort they put in or the volume of arrests they made.

But what happened, of course, was that the pressure associated with the high variability in results pushed back against the acceptance of the theory of value creation. The Regional Director in Kansas City complained that he had only class two or three violators in his region. He argued that the system "wasn't fair," because it didn't allow him to succeed. The right answer to that complaint was "Well, if you've only got class 2 or class 3 violators in Kansas City, you aren't a high productivity DEA region. Consequently, we ought to take resources away from you and put them in New York." But that meant organizational inequity, and it was hard for people to accept an untested theory of value creation against the desire to maintain fairness in the organization.

So, this measurement system collapsed, not because it wasn't useful but because there wasn't a powerful enough external drive to drive this conception through the organization and actually have it do the work of both directing and controlling as well as motivating performance and allocating resources.

One other example. I was working as a consultant to the Clark Foundation's Justice Program. Their goal was to use their limited resources to accomplish the goal of reducing "unnecessary incarceration." Needless to say, that was not a very popular goal in the states in which they were working. But the Foundation Board wanted to continue the program as long as it was working. What they wanted to know was whether it was working. My job was to help them make that judgment.

It turned out that the manager of the Justice Program had a theory that he was pursuing. But that theory was intuitive and implicit rather than explicit. Part of our job, then, was to dig this theory out and make it explicit. That turned out to be quite interesting. His idea was that the ultimate effect that interested him was the absolute number and share of offenders in a state who were, in his view, "unnecessarily incarcerated." He had an idea of what he meant by "unnecessarily incarcerated," and that was people in state penitentiaries who had committed only minor property offenses and had relatively minor prior records.

He thought the number of those unnecessarily incarcerated was the outcome of a process that started with: 1) public perceptions and political forces in the state that demanded more or less strict punishment (or official perception of these forces); 2) the policy networks that were active in shaping imprisonment policies; 3) the policies and practices that emerged from these policy networks (as they were affected by the reality or their perceptions of what citizens wanted); and 4) the existence of an infrastructure of alternatives to imprisonment in the state.

It was also true that the Foundation had some activities they supported that were focused on these different parts of the system. They conducted surveys of citizens to find out what citizens wanted and counteract widespread perceptions among politicians about what the people wanted. They helped to organize and sustain policy networks, and to feed information into them. They ran training programs for judges on sentencing difficult cases. And they supported the development of alternatives to incarceration.

With this logic model, it became possible to catalogue and measure the amount and character of activity the Foundation was producing with its limited resources, and so met some of the demands for accountability and performance. What remained difficult, however, was to measure the level of "unnecessary incarceration." That was obviously a value-laden concept, and there were intense conflicts between the Foundation and the states over how to define this result.

For a long time, we'd go around proposing different definitions. The conflict was so intense, however, that we'd never reach an agreement. Every time we got to a conceptual measurement of "unnecessary incarceration" somebody would say "that's the wrong definition."

Well, eventually we solved the problem. We did so not by agreeing on a particular definition of "unnecessary incarceration." We agreed, instead, to develop and report to one another in terms of a particular matrix that became the heart of the performance measurement system for the project. The matrix had the characteristic that it could be used to support different definitions of "unnecessary incarceration."

The basic idea was that, at a particular time, a state had a certain number of inmates. Those inmates could be distributed across a matrix like the one presented here as Figure 2. The rows of the matrix define the current offense for which the inmate is serving time -- ranging from murder through robbery, to larceny, to DWI and weapons offenses. The columns of the matrix define the prior record of the offender ranging from multiple prior felony convictions, through misdemeanor arrests, to no prior record.

Obviously the candidates for those to be considered "unnecessarily incarcerated" lie in the southeast quadrant of this matrix -- namely, those inmates with relatively less serious offenses and few prior convictions. Those who are candidates for "necessary incarceration," those inmates that no one would recommend releasing, lie in the northwest quadrant of the matrix.

What was wonderful about this matrix was that once we got the numbers in the matrices for the different states, people could draw their own lines around the parts of the matrix that constituted "unnecessary incarceration." Some would say our definition of incarceration is boxes two, three and six. Somebody else would say well, our definition of "unnecessary incarceration" is boxes eight, seven and twelve. The different definitions could still allow work to go forward in trying to reduce unnecessary incarceration even though people disagreed on the definition!

So this matrix could hold all particulars of various definitions of "unnecessary incarceration." We could get accurate information about what the true state of the world (namely, how many inmates were in each category). Then, each person could bring their own particular definition of unnecessary incarceration to determine how much of it there was. Even more interestingly, the matrix made it clear how different each state's view of unnecessary

incarceration was from the others. And that raised interesting questions about why one state defined it much more narrowly than another. For its part, the Foundation could apply its own definition, and on that basis answer the question of whether the program was succeeding or failing in their own terms. (Figure 2 sets out the basic logic model and the performance measures for the Clark Foundation's State Centered Program.)

VI. Conclusions

So, these are offered as illustrations of why there might be some leverage in the relatively simple activity of *performance measurement* compared with the more elaborate efforts of *program evaluation*. To suggest that such activities are simple is wrong, of course. They are simple only in a technical sense. Compared to the demands of program evaluation, performance measurement is analytically quite simple. It does not require measurements of outcomes. And it does not require reliable attribution.

What is hard about making a performance measurement system work, however, is not the technical pieces; it is all the political, organizational, and data system development work that has to go on to make it useful. To work over the question of what it is that we're trying to produce, and to get a clear quantitative picture of that at the level of output as well as at the level of outcome is really hard work inside an organization. Those in organizations resist doing the work with grim determination. But, in my view, it is the work that we must do if we are to have accountable, high performing public sector organizations. It is at least necessary for the work of program evaluation. And it might be sufficient to obtain many of the benefits that we thought were available only with a more (technically) rigorous program evaluation effort.

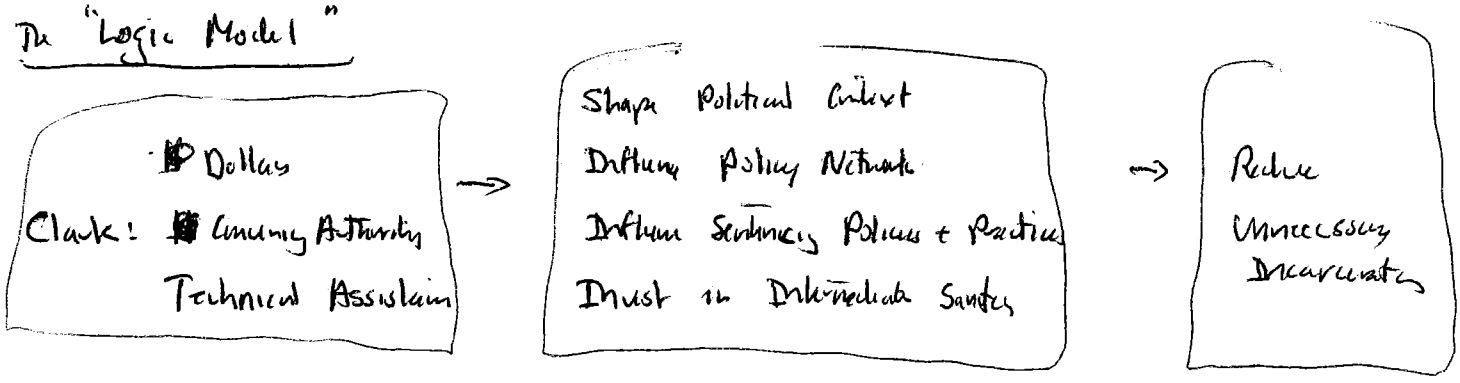
So, that's my argument. Let's start first with logic models, and measurement of processes and outputs rather than concentrating so much on the measurement of outcomes. It is technically easier and managerially more important to do so. Thank you for your attention and consideration.

****

Figure 2

The "Logic Model" + Performance Measures for the Clark Foundation's "State Centered Program" to Reduce "Unnecessary Incarceration"

I. The "Logic Model"



II. Measuring "Unnecessary Incarceration"

Current Offense	Multiple	Number of Prior Offenses: Few	None
Murder	Necessary Incarceration		
Robbery			
Larceny			Unnecessary Incarceration ?
DWI		Unnecessary Incarceration ?	Unnecessary Incarceration ?