Performance Evaluation as a Distinct Hybrid Form of Research Methods

Abstract

Performance evaluation should be viewed as an applied hybrid method distinct from program evaluation (defined by Peter Rossi) which combines methods from four different institutions each speaking a different policy language (Policy, Management, Programs and Accountability) as well as mixed methods. This includes two meta-data issues: Process Management ((the method of the "business process") includes the management using participatory evaluation of a process-based chain of events leading to desired outcome(s)) and Research Metrics (a solid grounding in the "facts" about whether there is progress towards desired goals). Performance evaluation provides new ways of defining reliability and validity and the three types of performance evaluation *inferences* are reviewed: 1) **relationships** – how are two outcomes linked?; 2) **outcomes** – how good are the indicators of outcomes?; and 3) a **process** – *what are the events and activities that produce outcomes?* Each factor is critical for a robust, functioning performance evaluation system.

Denise L. Baer, Ph.D.
Strategic Research Concepts
P.O Box 5729
Bethesda, MD 20824
PH: (240) 498-5409
Email: dbaer@strategicresearchconcepts.com

A paper presented at the American Evaluation Association Annual Conference, Washington, D.C. October 16-19, 2013

Introduction

Performance evaluation, now a global phenomenon and arguably also a "movement" in a growing array of governmental administrations at all levels, has been prominent on the national level in the U.S. for over twenty years – and yet it remains quite misunderstood from the perspective of research methodology. Consider this description of performance evaluation from the Organization for Economic Co-operation and Development (OECD), an organization developed after World War II then comprised of European countries, the U.S. and Canada dedicated to global development which emphasizes the information, policy and management aspects over the research design issues:

> The performance orientation of public managements is here to stay. It is essential for successful government. Societies are now too complex to be managed only by rules for input and process and a public spirited culture. The performance movement has increased formalized planning, reporting, and control across many governments. This has improved the information available to managers and policymakers (OECD 2005, p. 81).

In addition to admittedly being a distinctive policy and management tool, I argue that performance evaluation should also be understood as a research method in its own right. When understood as a method, performance evaluation joins but does not replace other research methods such as traditional qualitative (observation, document analysis, interviewing and focus groups) and quantitative (survey, experimental, large-N) methods or program evaluation. Just like these other methods, performance evaluation has its pros and cons – questions it answers better (or worse) than other methods. Unlike other methods, performance evaluation can work flexibly to evaluate strategies as well as the "3Ps" – policies, programs and projects. It is also uniquely applied and problem-focused because it focuses on populations at the macro level.

The risk of performance evaluation is that because it requires management and policymaker support and input to be successful, it is can be misused through mismanagement so that outcomes are not achievable or measurable. And because it promises public constituencies to provide data-driven policies, it can also be misused as purported "objective" criteria for achieving predetermined political purposes even where there is no data. And this is exactly the problem that national public policy currently faces. At the national level, the popularity of performance evaluation as a management tool has waxed and waned as different presidents have adopted different management styles and approaches to suit their own political purposes and policy strategies.

I turn first to an analysis of how different presidential initiatives have – while raising the profile of hoped for evidence-driven politics and effective programs – also misdirected our understanding of what performance evaluation is from a research methods standpoint. This background is important in advance of making a research methods argument for defining performance evaluation as a hybrid method. There are other, non-research paradigms, that have been offered to define how we should understand performance. To accomplish this, there has been an effort to subsume performance evaluation within an existing research methods paradigms – notably the positivist construction of the experimental method (touted as the "gold star" method). This approach ignores the possibility of methodological innovation contained within performance evaluation. Other alternate approaches which also blindside us to the methodological understanding include performance measurement, performance management, policy analysis, and program evaluation paradigms. I argue that performance evaluation, while borrowing from these paradigms and methods, should also stand alone, in part, because it offers two distinctive contributions to

more effective policies and programs – a focus on a macro-level outcomes and accountability. Finally, I explain how a research methods approach focused on research method generated criteria (e.g., causality and definitions of reliability, validity and accountability) can better address the existing tower of babel among our policy-making and implementing institutions as they strive towards evidence-based policies and programs.

The goal here is to lay the foundation for defining and creating a justifiable "commons" for valid and reliable performance evaluation that is shared by methodologists, program managers, and policy makers and donors. First, it is helpful to preview the advent of performance evaluation as a public policy.

The Waxing and Waning of Performance Evaluation as National Policy

At the federal level, performance policy development at the federal level began with the Clinton administration in 1993. That year saw passage of the Government Performance and Results Act (GPRA), a congressionally generated piece of legislation which required agencies to create multi-year strategic plans, annual performance plans, and annual performance reports. The goal was to get agencies to focus on results rather than inputs or operating procedures using their defined missions. GPRA initiated a government-wide agency-level reorganization as well as providing for new congressional oversight and reporting requirements over performance. This new law was paralleled in the executive branch by new executive branch initiatives led by Vice President Al Gore which encouraged agencies to define indicators of their missions. While there was an effort to point to a reduction of redtape and a simplification of cross-agency duplication of "processes" and "programs," the overall effort stress the measurement side of performance. For President Clinton, the political benefit was to present himself as a new type of Democrat who was able to downsize and reduce government.

Congress failed in the 1990s to create new structures to support its newly undertaken performance evaluation responsibilities. In 2001, this enabled the newly elected President George Bush (Bush 43) to claim presidential preeminence over Congress on agency performances. President (Bush 43) redefined performance measurement as a presidential and Office of Management and Budget (OMB) management tool (rather than a joint policy process with Congress), consistent with his general philosophy of acting without Congress using signing statements and presidential orders and agreements. The emphasis here became the management side of performance. OMB created a Performance Assessment Rating Tool (PART) to grade federal programs on an ineffective-to-effective scale based upon four criteria (program purpose and design, strategic planning, program management, and program results/accountability). PART scores about effectiveness were then used to make presidential budget decisions, often without investing in evaluation or good data. For Bush 43, the political benefit was to assert control over government management.

In 2009 at his inauguration, President Barack Obama proclaimed his commitment to data-driven policies saying "The question we ask today is not whether our government is too big or too small, but whether it works – whether it helps families find jobs at a decent wage, care they can afford, a retirement that is dignified. Where the answer is yes, we intend to move forward. Where the answer is no, programs will end." Formally, deciding what works for Obama at the presidential level has included a move from overall agency performance alongside a new emphasis on use of randomized control trials (RCTs) (the Program Evaluation Initiative) for selected large projects based upon directives from the Office of Management and Budget designed to increase scientific rigor. Overall formal performance evaluation has been relegated to

the primary purview of cabinet secretaries based upon agency priorities (the High Priority Performance Goals (HPPG) Initiative) rather than something Obama managed at the presidential level. This restructuring of performance requirements was institutionalized with passage of the Government Performance and Results Modernization Act (GPRMA) in 2010. GPRMA revised agency strategic planning and performance planning and reporting requirements with an emphasis on cross-cutting federal priority goals and agency-level priority goals. GPRMA also legislatively authorized agency level chief operating and program improvement officers, a government wide performance improvement council and a performance website. However, in informal terms, the Obama administration has also implemented the use of policy "czars," up to as many as 45 including one being a "performance" czar, with another 18 announced but unfilled positions, whose responsibility it is to manage discrete policy areas across federal agencies (Judicial Watch, 2011). While some of these policy czars have departmental appointments while others are presidential advisors, they report directly to the White House and are largely unaccountable to Congress, and thus also lack public performance accountability outside of internal decisions of the Obama White House. When both the formal and informal approaches of the Obama administration are considered together, the executive branch has reduced its oversight of performance to items of major campaign agenda or constituency concern, unless it is an appropriate subject for RCTs. The political benefit for Obama is to politicize his control of his agenda, while claiming to be more "scientific."

Another factor that has accelerated the overall trend of decreased attention to performance and data-driven policies has been the dominance of budget cuts on the policy agenda. This is a trend that started in the 1990s for reasons unconnected with the performance movement. But following the budget downturn in 2008, cuts have been made to programs not based on evidence about their effectiveness, but based upon their cost. Using econometrics, and across-the-board budget cuts to programs and agencies (sometimes done by agreement and sometimes done via sequestration) regardless of their ability to achieve outcomes, program outcomes are less regularly discussed on the national news. Two prominent exceptions are policy changes being considered in how the government and the private sector reimburses physicians (whether this should be fee-for-service or based on patient or other health outcomes?) and teachers (should it be solely mandated testing or some other way to assess student outcomes?). Both of these singular prominent policy agenda items emphasize personnel and individual accountability for results that than program or policy accountability which had been present in the Clinton and Bush 43 administrations. Percentage cuts fail to address whether there is a critical minimum level of funding that provides a reasonable opportunity to achieve outcomes. They also fail to incorporate performance and data-driven information about outcomes with their emphasis on program costs. In these as well as in other arenas, however, performance evaluation remains vitally important for both the government and the nonprofit community, and increasingly important in how many of us at the individual level are evaluated in yearly performance reviews.

Why It Matters to Define Performance Evaluation

I turn next to what's in a name – the vexing problem of terminology which I argue is part of the fog that has obscured performance evaluation as a method. There are four "fogs" that I discuss that are also terminology issues:

- Academic disciplines' "separate tables" selective treatment of mono- vs. mixed research methods;

- The political interests of "Big Science" and the growth of Randomized Control Trials (RCTs) and large-N methods in establishing federal priorities;
- Measurement vs. management vs. evaluation terminologies; and
- Policy analysis vs. program evaluation vs. performance evaluation distinctions.

What's in a Name I?
How the "Separate Tables" Debate has Blinded Research Methodologists to Key Methodological Issues in Performance Evaluation

While it is certainly true that the practitioner community has responded to the performance movement, the silence from the more academic methodology community has been deafening. Most researchers learn their research methods within their discipline of training, and so it is not surprising that these disciplinary lenses color the types of methods one is trained in even though the actual epistemology of methods is considered universal. I was trained as a political scientist and use this discipline as an example of this limited purview of the scope of methods. By the universal epistemology, I mean that there are common philosophical terms for understanding research methods, and not the positivism position of the unity of all sciences. Viewed in this light, there are three methodological communities (quantitative, qualitative and mixed) which dominate research method discussions (Teddie and Tashakkori 2009).

TABLE 1
Separate Tables:
Three Methodological Communities

Three Communities

Qualitative Methods *stress "external" validity, context-rich narratives and idiographic explanations.*
- Interviewing
- Field Work
- Participant Observation
- Focus Groups
- Traditional Case Study
- Causal Case Studies

Quantitative Methods *stress "internal" validity, statistical analysis to demonstrate non-spuriousness, and nomothetic explanations*
- True and Quasi-Experiments
- Surveys
- Observation
- Large N-Designs

Mixed Methods *stress "meta-inferences" based on triangulation and complementarity of data from both quantitative and qualitative methods*
- Parallel Mixed
- Sequential Mixed
- Conversion Mixed
- Multilevel Mixed
- Fully Integrated Mixed

However, it is a fair statement to also say that within academia, the methodological communities of qualitative and quantitative approaches comprise not just different types, but actual "separate tables" compared to mixed methods. Gabriel Almond (1988) has used this metaphor to describe how political scientists rely on discrete methods tables to support their lonely and unhealthy commitment to particularistic ideological and theoretical positions. Almond cites the play "Separate Tables," which was a "hit of the 1955 New York theatrical season [where], the Irish playwright, Terence Rattigan, used the metaphor of solitary diners in a second-rate residential hotel in Cornwall to convey the loneliness of the human condition" (1988:828).

This metaphor can also describe divides between the social sciences. The discipline of psychology

through the work of Donald Campbell and his associates has played a central role in establishing randomized control design as well as quasi-experimental methods (Dehue 2001). The disciplines of sociology and anthropology (as well as community psychology) have depended more on qualitative methods, including field work, observation, case studies, process-tracing, narratives and story-telling. It was the field of sociology, rather than public administration, which largely gave birth to program evaluation through the work of Peter H. Rossi and his collaborators, Howard Freeman (a sociologist) and later, Mark Lipsey, a psychologist.

Political science, by contrast, historically has been quite eclectic and has borrowed freely from economics, sociology, psychology, history and the law and has included all types of research methods (Morton and Williams, 2010; Sartori, 1988). In recent years, however, the current debate within political science has increasingly pitched quantitative methods (King, Keohane and Verba, 1996) against qualitative methods (Brady and Collier, 2010). The quantitative camp argues that while there may be two "styles" of research, there is a single logic of inference and a unity of the scientific method (King, Keohane and Verba, 1996). And each camp provides separate training – there is advanced quantitative training at the University of Michigan, while there is advanced qualitative training at Syracuse University.

But the dominant paradigm remains quantitative, which lends greater credence to the experimental method underlying RCTs as a value within political science methods training. For example, an analysis of political science research methods found "a consensus on positivism as the mode of scientific research in political (and social) science, either stated explicitly or implied through various structural and rhetorical devices [with the message that]..."empirical research" and/or "the best research" is quantitative research" (Schwarz-Shea and Yanow, 2002:476)

Within political science, performance measurement has been primarily studied by the governance and public administration specialists more interested in management issues, not by the methodologists (to be also considered in subsequent sections). For example, a search by the author of JSTOR using the search terms '"performance measurement" and evaluation and validity and reliability and method' came up with a total of 475 articles from the 1960s to the present. The journals did not include premier disciplinary associational journals, and none included a major consideration of research methods criteria and epistemology. Instead, the journals represented included specific topics related to employee appraisal or outcome policy areas (e.g., public health, transportation, education) or those covering management issues coming from business, public administration and management.

The third community is mixed methods are only about twenty years old as a formal methodology (Johnson, Onwuegbuzie and Turner, 2007)., although as noted above, not all research methodologist agree with the premise that there are multiple methods. While program evaluators typically use mixed methods, program evaluation as a methodology is somewhat older than mixed methods approaches. As a method, mixed-method researches have due to growing acceptance of multiple ways to measure causality (Cresswell and Plano Clark, 2007; Tashakkori and Teddlie, 2003; Brady and Collier, 2004). Whereas each stand-alone method has a specific trade-off in terms of internal and external validity, use of the mixed-method/multi-models approach allows for enhanced causality due to triangulation (where results of one type of method are measured against another (Bryman 2006), the ability to offset the weakness of one method against another, and completeness, where a more complete answer can be found. This mixed methods approach rejects the idea that each method is a separate table or paradigm that is an incommensurable paradigm –

instead, data are seamless, and more complex designs permit the collection of data that can be multi-purpose.

It is important to note that program evaluation is also known as a hybrid method because in addition to using multiple methods, they apply them with specific evaluator competencies and skills. Program evaluation was created with precisely this argument (Langbein and Felbinger 2006). Rossi and Freeman suggests that what is unique about program evaluation the "fitting evaluations to programs" (Rossi & Freeman, 1985, p. 102). Carol Weiss argues that evaluation is a hybrid because it is done in a political context: "Politics intrudes on program evaluation in three ways: (1) programs are created and maintained by political forces; (2) higher echelons of government, which make decisions about programs, are embedded in politics; and (3) the very act of evaluation has political connotations." (Weiss, 1991: 213). Other evaluation approaches on the evaluation theory tree include considering stakeholders and decisionmakers, as well as participatory evaluation methods (Akin and Christie, 2004).

Academic research methods has remained relatively silent on performance evaluation as a method and we turn next to the broader question of whether performance evaluation can be subsumed within RCTs?.

What's In a Name II?
Can Performance Evaluation be Subsumed Within "Big Science" RCTs?

"Big Science" is a term that was used to describe program evaluation as it emerged in the 1960s. While the term "big science" originated in the post World War II era to signify large government investments in research laboratories and research teams managed by large universities as opposed to research by individuals, in the social sciences, big science refers to regular investment in randomized control trials focused on impact. For example, one major independent review commissioned by the U.S. Congress of all crime prevention programs funded by the U.S. Department of Justice (DOJ) concluded that "big science" and impact evaluations provide the only valuable type of research:

> spending adequate funds for strong evaluations in a few sites is far more cost-effective than spending little amounts of money for weak evaluations in thousands of sites. ...Evaluation funds should be conserved for impact assessments. ...enough funding for strong science. Such studies routinely cost $15 million or more in other agencies and are often mandated by Congress, but there is no precedent for such "big science" at DOJ (Sherman et. al., 1998:13).

In contrast, what has thus far constituted performance measurement methods and management approaches came out of the practitioner and not the research community. This is an important point in understanding the delayed reaction of "big science" to performance measurement and management coming almost a decade after its introduction. This is true whether one dates the performance orientation to the 1990s, or traces it to earlier efforts in the international development or domestic social service communities (all of which will be discussed in depth later). In the 1990s, it is noteworthy that the anointed proponents and early adopters of the new performance movement were practitioners came from outside of government as well as an ambitious presidential candidate (Bll Clinton) who defeated a competitive field of aspirants in his own party as well as an incumbent president (Bush 41). According to David Osborne and Ted Gaebler, the authors of *Reinventing Government*, an instant *New York Times* bestseller in 1992, the

focus on performance was designed to map out "a radically new way of doing business in the public sector" (Osborne and Gaebler, 1992: xviii). Anointing of the Osborne and Gaebler book came from the very top. Presidential candidate Clinton, who made reinvention a centerpiece of his presidency when elected in 1992, is quoted on the book jacket saying that "This book should be read by every elected official in America. Those of us who want to revitalize government . . . have to reinvent it. This book gives us the blueprint." President Clinton proclaimed in his first State of the Union speech that "We must reinvent government to make it work again" and assigned Vice President Al Gore to lead the federal government's reform efforts under the auspices of an interagency task force, the National Performance Review (NPR). At the outset, this was both an executive and a congressional-level revolution in the policy of evaluation that came outside of the existing evaluation "big science" community.

Following passage of the landmark 1993 GPRA Act which created new mandates for federal agencies, the sudden reset of national evaluation policy dollars from program evaluation to government performance demanded required new skills for evaluators and federal agencies and managers to work with a new set of research and program tools. For example, this included sets of opposing terms: outcomes (rather than outputs or impact), logic models (as opposed to program evaluability assessments), and using numbers to change the structure of programs (e.g., the concepts of statistical vs. process redesign or reengineering). These terms were at the time equally unfamiliar to agency staff familiar with producing programmatic outputs based upon program authorizations, and evaluators who had focused on programmatic impact using program evaluation methodologies..

The advent of performance evaluation in the 1990s produced a changed policy environment so sudden that it was the policy equivalent of a tectonic shift in how policy is made and evaluated. Just as geologists describe how the slow movement of layers of the earth can suddenly produce seismic results when the plates converge, diverge or transform their boundaries, the sudden tectonic arrival of performance on the national policy scene changed policy boundaries. While the new focus on performance may be linked to gradual trends in governance and public expectations as well as innovative state and local initiatives and even global precursors (e.g., Great Britain, Netherlands, Australia and others) that started as early as the 1970s and 1980s, the policy adoption at the national level in the U.S. in 1993 was sudden. Since then, scholars who map such changes have been scrambling to explain the large policy and research consequences with some acting as champions while others decry the performance movement and still others taking a more neutral role.

The novelty of the new approach to both research and management created an opening for those new to government consulting as well as new opportunities for existing professional services firms to expand (or maintain) their market share. Among older organizations which capitalized on the performance movement, notable is the Urban Institute (UI) which was organized in 1968. UI's Director of Public Management Harry Hatry became an instant expert on performance measurement and his book, *Performance Measurement*, which came out in 1999, was marketed as synthesizing "more than two decades of [Urban Institute] work into a comprehensive guide to performance measurement."

New organizations were formed also by existing consulting groups. Notable in this category is the IBM Center for the Business of Government which was organized in 1998 by the PriceWaterhouseCoopers, one of the world's largest professional services and accounting firms, and renamed when acquired in 2002 by IBM.

Yet others were enterprising entrepreneurs new to the field. This includes The Performance Institute, organized by Carl DeMaio, then an entrepreneurial young staffer with the Republican Party affiliated Congressional Institute, who saw a profitable business opportunity in organizing government training conferences for federal agencies. The Performance Institute, which now provides hundreds of trainings a year to federal and state employees, has also earned profitable contracts with federal agencies assisting them in meeting long-range federal planning requirements. DeMaio sold The Performance Institute in 2007 to the Thompson Publishing Group, a prominent consulting company providing regulatory compliance services.

An array of other groups offer commercially branded consulting and training approaches (e.g., The Balanced Scorecard Institute's "Nine Steps to Success"® or Motorola's the Lean and Six Sigma ® approach to process management). Yet other groups offer specialized training in the use of specific tools such as Key Performance Indicators (kpi's) or "dashboards" for management and this is a common function offered by an array of consultant firms and solo practitioners. The dominance of proprietary approaches was common in the field as it developed in the 1990s (Kettinger, Teng and Guha, 1997).

Large think tanks, consulting firms and universities who benefit from what had come to be known as "big science" have stressed what they call the "gold standard" of evidence-based policies – use of the experimental method rather than performance measures. The idea of the "experimenting society" has been around for some time. The father and theorist of the methods of experiments, Donald Campbell, articulated his vision of evidence-based society in 1969 where he advocated government investment in the creation of social indicators and data banks:

> The United States and other modern nations should be ready for an experimental approach to social reform, an approach in which we try out new programs designed to cure specific problems, in which we learn whether or not these programs are effective, and in which we retain, imitate, modify or discard them on the basis of their apparent effectiveness on the multiple imperfect criteria available. (Campbell, 1969: 409).

Experimental methodologies are valued because they provide the strongest set of research designs to establish internal validity. Experiments based on the principle of randomization with a control group (hence the name Randomized Control Trial or RCT) are based upon John Stuart Mills principle of method of difference. Presumably, if all individual units are randomly assigned to both the treatment and the control group, then the only "difference" is the cause. As such, they are able to measure what is typically termed "impact" of the "cause" which became to be distinguished from "outcome" (the term of choice in performance evaluation) as the effect where the amount of other causes (and error) are subtracted. The

idea of emphasizing strong evidence of effectiveness is something advocated by the National Academy of Sciences which is a congressionally chartered association charged with providing independent, objective advice to the nation on matters related to science and technology. In part reflecting their dominant membership from the physical sciences, medicine and engineering, NAS has produced reports arguing that rigorous evaluation through RCTs provides the strongest evidence of internal validity. For specific, discrete treatments or interventions, if it is possible to randomize treatments using multiple trials (or places or over time) and use controls (which randomly do not receive the intervention being tested), this will provide very strong and unparalleled design-based evidence about what can work and how much of an impact that treatment considered alone can have if all other factors are held constant.

This is something that performance evaluation cannot do – it cannot tell you the amount or size of the impact of a policy, program or project. But RCTs also have weaknesses since they are limited as to what they can test at any one time (how many "X's") and in their ability to take account of context (or external validity) for which a variety of other methodologies do much better at. RCTs are not a one-size-fits-all or the only model for establishing outcomes in all circumstances. Further, it requires often hundreds of repeated experiments to ensure adequate external validity and the weakness of RCTs is that use of them require others (e.g. What Works and the Campbell Collaboration) who review studies and determine their level of evidence using meta-analyses (Cartwright and Hardie, 2012:58). Yet, research interest and advocacy groups have become organized to argue that they provide exactly this type of model that should dominate how performance is examined by and invested in by the federal government. The types of grants or contracts

BOX 1
National Academy of Sciences (NAS)
Recommendation on Criteria for Establishing Strong Evidence of Effectiveness

Federal and state agencies should prioritize the use of evidence-based programs and promote the rigorous evaluation of prevention and promotion programs in a variety of settings in order to increase the knowledge base of what works, for whom, and under what conditions. The definition of evidence-based should be determined by applying established scientific criteria. In applying scientific criteria, the agencies should consider the following standards:

- Evidence for efficacy or effectiveness of prevention and promotion programs should be based on designs that provide significant confidence in the results. The highest level of confidence is provided by multiple, well-conducted randomized experimental trials, and their combined inferences should be used in most cases. Single trials that randomize individuals, places (e.g. schools), or time (e.g., wait-list or some times-series designs), can all contribute to this type of strong evidence for examining intervention impact.
- When evaluations with such experimental designs are not available, evidence for efficacy or effectiveness cannot be considered definitive, even if based on the next strongest designs, including those with at least one matched comparison. Designs that have no control group (e.g., pre-post comparisons) are even weaker.
- Programs that have widespread community support as meeting community needs should be subject to experimental evaluations before being considered evidence-based.
- Priority should be given to programs with evidence of effectiveness in real-world environments, reasonable cost, and manuals or other materials available to guide implementation with a high level of fidelity.

National Research Council and Institute of Medicine. (2009). Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities. Committee on Prevention of Mental Disorders and Substance Abuse Among Children, Youth and Young Adults: Research Advances and Promising Interventions. Mary Ellen O'Connell, Thomas Boat, and Kenneth E. Warner, Editors. Board on Children, Youth, and Families, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. The full report is posted at http://www.nap.edu/catalog.php?record_id=12480. The above is recommendation 12-4 in the report, on page 371.

that pay for RCTs are extremely attractive to large universities or firms which seek government funds to cover their infrastructure and staff, especially as foundations are donating less in the way of operating funds to think tanks and the states are providing less and less to public universities. An example of the type of group that represents these interests is The Coalition for Evidence-Based Policy, organized in 2001 by Jon Baron as a philanthropic organization to stress the role of randomized control trials in producing rigorous evidence. Interest groups like these have a stake in ignoring the challenge to a research methods classification of performance evaluation. We now turn to a discussion of measurement vs. management vs. evaluation.

What's in a Name III? Is it Performance Measurement, Performance Management or Performance Evaluation?

While some evaluators use all three of these terms (measurement, management and evaluation) interchangeably, each term references a distinct set of practices, each of which parallels the various presidential emphases on measures versus management. My argument is that the terms measurement and management are both limited conceptually, and do not reflect the criteria standard in research methods and science.

Performance Measurement. The Clinton/Gore performance approach developed a hypertrophy of specialized data bases within agencies. The idea was to identify and develop all possible indicators for international, state and local projects to choose from. From a scholarly perspective, the problem with understanding evaluation as measurement is that measurement is just one part of the scientific method. In the public and social spheres (the province of social sciences), measurement is uniquely derivative rather than determinative of program theory, program goals, and desired policy outcomes. Among laypersons, it may seem that measurement precedes science. In the history of the physical sciences, for example, measurement and agreement of what is being measured (e.g., temperature, time, distance, size, weight) actually preceded more advanced findings and theory-building. Thus, the history of science can be seen as the development of thermometers, microscopes, telescopes, scales and chronometers which allowed precise measurements to serve as the spur to theory. It is of course true that more recent developments in physics refute this idea. Usually stated in research methods in the form of the Heisenberg Principle, the newer understanding treats measurement as derivative because of the finding that the position and momentum of a particle cannot be precisely determined not only because of observer effects, but also due to what is recognized as a fundamental property of quantum systems (once you measure it, it changes what you are measuring). Less recognized among laypersons is that as the social sciences developed, the reverse is true: measurement clearly varies considerably based on the theory being tested and the multitude of causes in the field (laboratory effects are always stronger) as well as reflecting changes occurring solely due to the placebo effects of observation and testing (the purpose of control groups). This is why qualitative methods approaches have developed unique to the social arenas because of the difficulty of separating cause from effect when there are multiple levels of possible effects (e.g., individual, group, community, etc.). This is also why managing for performance cannot be simply relegated to those with technical skills or used to working with a specific dataset, or assigned to a low-level staffer to monitor. Outcomes cannot be measured in a hypothetical "technical" vacuum. Those who make policies and decide goals, those who administer and manage programs, and those whose job it is to devise performance

measurement approaches and systems must all work together to produce the desired result. This requires that all understand the methods of performance evaluation, even though each plays a distinctive and different role in the policy and management process of delivering outcomes.

Performance Management. Both the Bush 43 and the Obama administrations have placed greater emphasis on management of performance, with selective focus on defined programs and projects within agencies to agency priorities and free-ranging White House "czars." In part, this came as a result of criticisms of the Clinton era where too heavy a focus on indicators and unmeasurable outcomes resulted in a degradation of management when we depend on managers and experts to provide "specialized and technical knowledge" critical to delivering a service or implementing a public policy (Radin, 2006:2).

As James Thompson has argued, "Measures can greatly enhance the management function; they also can serve as an agent by which that function is eroded. It is substantially in the hands of top management to decide which approach is taken. The decision is complicated by the political dimensions of measures; the potentially centralizing effect of measures can directly enhance the power of top management and provide an improved capacity to respond to the demands of key stakeholders. There is a need, however, to balance these short-term considerations with the longer term need to maintain and develop the organization's overall capacity to manage and adapt. " (2000: 280).

From a scientific research methods standpoint, management beyond research administration is admittedly not well-understood. Typically, a research project may require project administration, but it is management solely of the research project. Within the management field, managing for performance is just one subset of management skills. Managers are trained differently than researchers and have an entire portfolio of other sectors to manage besides research – budget, human resources, vendors and procurement, stakeholders (e.g., boards, partners, clients, customers and other constituencies) and public relations as well as programs and events. These are separate skills from research.

To assist managers, Increasingly, performance management approaches include the use of management "dashboards" (a graphic summary of key indicators being tracked for management oversight) or an emphasis on tools such as "monitoring and evaluation" (M&E) used throughout the life of a specific project to ensure that it is managed to produce an evaluable project and tailored to produce the desired outcomes. Another new approach is to create a nimble agency or organization (known in the field as a knowledge or learning organization) that is able to quickly respond to changes in the environment or to changes in how well their work is received or progresses. As a result, many refer to performance management or more awkwardly as "outcomes-based performance monitoring and evaluation (M&E)" rather than performance measurement to subsume this awkward marriage. M&E combines management monitoring as a continuous process with how policies, programs and projects are evaluated. The World Bank, for example, describes M&E as an overall governmental system that once institutionalized produces better policymaking and administration (Mackay, 2007). Others view performance as part of governance.

> Building an M&E system essentially adds the fourth leg to the governance chair. Typically and traditionally, governments have built budget systems, human resource systems, and auditing systems as the three legs of a governance system. But what has been missing has been the feedback system on the outcomes and consequences of government actions. This is what building an M&E system brings as an additional public sector management tool. Kusek, Rist and White, 2004: 3

http://www.ideas-int.org/documents/file_list.cfm?DocsSubCatID=13

This dominant management approach increasingly treats performance management as one of an array of management processes including financial, human resource and strategic management (Halligan, 2001). Those who emphasize management over methods tend to stress that it is the job of managers rather than researchers or analysts to selected performance indicators. For example, Harvard Kennedy School's Robert Behn argues that "When the top officials in any governmental jurisdiction or public agency delegate to their analysts the task of selecting the organization's key performance measures, they are ducking their leadership responsibility" (2011: 2). This approach is aligned with new approaches to management known as the New Public Management (NPM). NPM challenges traditional public administration by emphasizing market forces and manager freedom to manage rather than working with established procedures and legally defined parameters. Within NPM, performance became more of a management purview rather than a research tool. The problem is that aligning performance evaluation with management as if they are one and the same creates a management philosophy rather than clarifying what performance evaluation actually is so that the concrete steps in how be to be successful in doing it effectively are lost or minimized.

Developing management competencies for using performance data is, of course, an important area. As noted above, one of the criticism of GPRA was that it focused too much attention on collecting data on too many indicators rather than how to use the information as an effective management tool. The problem is that in obscuring the pros and cons of diverse methods, the danger is failure to ensure valid and reliable methods and thus results. That is the *sine qua non* of science and scientific research methods.

We now consider whether performance evaluation is a subset of policy analysis or program evaluation.

What's In a Name IV? Is it Policy Analysis, Program Evaluation or Performance Evaluation?

Policy analysis, program evaluation and performance evaluation all promise to provide evaluation. They differ in terms of what is evaluated as well as the methods used, and to argue that performance evaluation stands on its own as a distinctive method requires distinguishing performance evaluation from each (see Table 2).

Policy analysis focuses on a policy area or problem, which might group together a variety of public programs as well as examining problems and non-decisions that are not covered by extant policy. While different disciplines may have different approaches to the study of policy, typically all do employ a range of conceptual and analytical tools such as policy space/policy maps/political feasibility assessments, public interest, stakeholder analysis, agenda setting, policy framing, issue networks and iron triangles, issue evolution, policy cycles, and decision-making (Bachrach and Baratz, 1963; Brewer, 1974; Dery, 1984; Baumgartner and Jones, 1993; Stokey and Zeckhauser, 1978; Heineman, Bluhm, Peterson and Kearney, 1990; Burstein, 1991; Howlett and Ramesh, 2003; Smith and Larimer, 2009). Policy analysis as evidenced in the growth of both schools of public policy in the past twenty years as well as program evaluation are two different analytical and research approaches that provide information to policy makers

just as performance evaluation purports to do. Policy analysis (at least within political science) is dominated by the dominant quantitative and positivist paradigms. For example, Paul Sabatier's comprehensive survey of policy analysis theories (stages heuristic, institutional rational choice, multiple streams, punctuated equilibrium, advocacy coalitions, policy diffusion, large-n comparative studies of the funnel of causality) is based upon the selection criteria of following "scientific norms of clarity, hypothesis-testing [and] acknowledgment of uncertainty" Sabatier, 2007:11).

Program evaluation, by contrast, typically focuses on programs. As such, it requires an evaluable program. As Rossi, Lipsey and Freeman conclude, program evaluation is only appropriate for "mature, stable programs with a well-defined program model and a clear use for the results that justifies the effort required" (2004: 59). Much of the growth of new areas of program evaluation has occurred with new programs and programs as they are being developed and include participatory evaluation as well as formative evaluation methods which use evaluator skills to help define an evaluable program.

TABLE 2
Public Policy Evaluation Methods

COMPARISON CRITERIA	METHOD		
	Program Evaluation	**Performance Evaluation**	**Policy Analysis**
Major Questions Addressed	What Impact Did the Program Produce? Was the Program Implemented with Fidelity	Do Programs Lead to Desired Outcomes? Does the Program Need Adjustment to Work Better?	Are Public Policy Strategies, Policies and Programs Aligned with Current and Future Policy Problem? What is the Policy Space? Are There Policy Gaps and Policy Silences?
Empirical Focus	Programs	Programs Policies Strategies Management Functions	Policies
Theoretical Focus	Program Theory	Business Process	Policy Area or Problem
Unit of Analysis	Program Participants	Accountable Entities Population-based Measures	Policy Problems Policies
Who Conducts Research?	Usually EXTERNAL Evaluator	Usually INTERNAL Evaluators	Can be either INTERNAL or EXTERNAL
Time Perspective	Discrete Time Frame Can Require up to 5 Years or More for Full Results Backward Facing *(unless participatory approach)*	Continuous Time Frame Weekly, Monthly, Yearly Trends Backward and Forward Looking	Discrete Time Frame Can Be Extensive Backward and Forward Looking
When Results are Available	Results Available After Evaluation Complete May Take 5 Years or More	Immediate Continuous Links Before and After Methods	As Needed Can Address Current and Future Problems

Management Feedback Loop	Lengthy – results may take 5 years or more	Dashboard Approach Immediate Intended as 1 year or less	Flexible Method
Evaluators Skills Needed	Knowledge of Policy Area Basic Research Methods Experimental or Large-N Methods	Knowledge of Policy Area Basic Research Methods Social Indicators Management Business Process	Knowledge of Policy Area Can Be Qualitative or Quantitative

Performance evaluation, contrast, differs in that it focuses on accountable entities and the business process as well as providing a longer time frame. While process analysis often involves proprietary techniques, analyzing processes involve distinct steps and include diverse multi-methods approaches (Kettinger, Teng, and Guha, 1997).

A process is distinguishable from a program and a policy as well as a "cause". A process can be defined as a series of activities and outcomes that occur in sequential order in order to produce an outcome. These steps are usually present graphically in the form of a logic model with boxes and arrows demonstrating how the steps occur over time. If processes are not managed well, they can produce counts of activities (outputs), but fail to produce the outcome. For example, a large number of students can engage in service learning (the count of students providing a set amount of volunteer hours of community service), but fail to achieve the character change intended or provide any real benefit to the community. Processes can be managed for incremental (or statistical) improvement to reduce "cycle time" between steps, or redesigned (eliminating or adding steps) to radically increase effectiveness and the achievement of outcomes. Most programs comprise one or more processes. The business and particularly the information technology (IT) communities made a distinctive contribution to the concept of performance in introducing the ideas of the business process, process management and outcomes. However, other aspects of change interventions such as comprehensive community initiatives and the development community can be understood as a "business process." It is in these arenas that logic models and logframe analysis developed, for example. Much of what occurs that is evaluable can be defined in the nature of a process. And it is this new methodological innovation that remains unappreciated. Let's now consider how performance evaluation provides different products in evaluation.

Distinctive Methodological Aspects of Performance Evaluation

Like program evaluation, performance evaluation is a hybrid approach, but for different reasons. Performance evaluation simultaneously:

- focuses on **performance** – *which matters to both managers and policymakers – but for different purposes,* and.
- focuses on **measurement** – *which has very different meanings in the business environment and within science.*

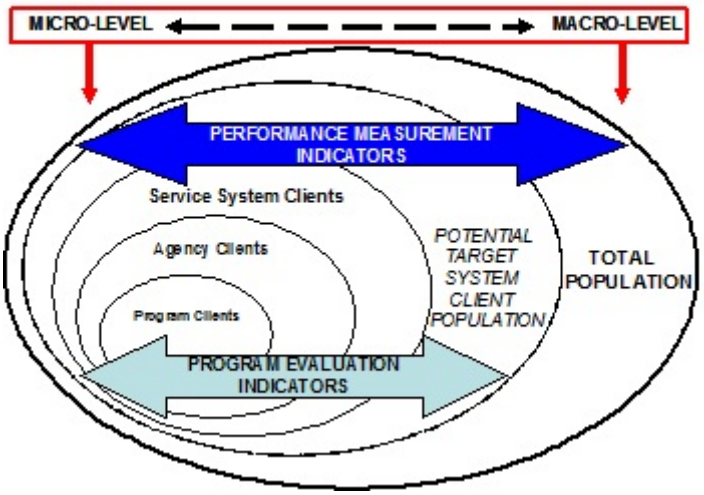
Compared to other research methods, I argue that performance evaluation is distinctive because it promises:

- To be able to measure how well public programs and policies work by measuring outcomes that bridge the micro-macro policy gap.
- To ensure that government is accountable for its programs and policies – a critical factor for democratic government.

What makes performance measurement a distinct method is the fact that it seeks to fulfill both promises. Yet, these promises are extravagant in that performance measurement is not the only research method which makes these promises, and performance measurement alone cannot deliver on these promises. Let us examine each promise in turn.

Promise #1: Measuring Outcomes and How Research Methods Address the Micro-Macro Gap

Figure 1: BRIDGING THE MICRO-MACRO GAP



Performance measurement promises to focus on the micro-macro gap. The micro-macro gap (or paradox occurs when program evaluations demonstrate the effectiveness of individual projects or programs, but there continues to be a failure in overall improvements in the conditions or the social problem that the programs seek to address. From a government perspective, this promise is new because it moves the attention of the government bureaucracy toward the effects of programs rather than the procedural (e.g., output) or legal aspects of how it is implemented. From a research perspective, this promise is new because it reflects a new methodological focus on a new type of macro-level data – outcome data. Outcome data alone promise to address the micro-macro methodological gap.

Outcomes are measured by social indicators and reflect aggregate or macro-level data on populations. In contrast, micro-level data is individual-level data collected only on program or project participants. Since programs and policies are rarely "full coverage" where all in the population are participants, performance measurement outcomes in their purest form are "gross" outcomes. There are, of course, ways to use performance measures at different stages of a project, program or policy – but what distinguishes performance measures is that they are always linked to macro-level data. This requires specialized performance evaluation measurement methods. Impact results are different – they are "net" outcomes where an impact measures the results after other and confounding effects on the result have been removed. Typically, impact results are tied to only those who received a program or intervention and not the broader population. Typically, impact results are measured using program evaluation methods as well as RCTs and traditional mono-methods.

As shown in Figure 1, there are several distinct client groups or populations to which one might wish to attribute outcomes: project or program clients, agency clients, service system clients, the potential target

system client population, or the total population. Social indicators – the focus of performance measurement – define the total population. In contrast, program evaluation indicators are not typically population-based, although one can calculate intended and actual levels of coverage for larger population groups. Typically, program evaluation focuses on measuring causal impact, but only for those who participated in the program.

Performance indicators can be used at the program levels, but they typically require that all measures be linked to or target outcomes. It is only performance measurement indicators that aim to provide measures of total population outcomes. Having good measures that can provide actionable information about macro-level outcomes is new. The only other research method that potentially promises to target macro-level data are "Large-N" or econometric methods. Econometrics is an established method that also seeks to measure macro-outcomes. But econometrics has thus far failed to make the link in ways that satisfy a broad range of methodologists. In one view, for example, econometrics utilizes empirical models based upon regression modeling that "relies on ad-hoc specifications that have little or no theoretical basis," and are thus susceptible to model specifications and technical data problems, such as "parameter heterogeneity and endogenous regressors, measurement errors, influential observations and error correlation" (Quibria, 2004: 17).

No other method currently satisfactorily addresses the micro-macro paradox. Some methodologists such as Hubert Blalock have gone so far as to argue that the paradox is unsolvable – due to the ecological fallacy. Building on the paper by W.S Robinson (1950), Blalock discusses the potential philosophical problem as we shift units of analysis – moving from individuals to larger units of aggregates. Blalock argues that the key to the problem is that "in shifting units we may be affecting the degree to which other unknown or unmeasured variables are influencing the picture (99).

Consider, for example, USAID, a federal agency which was an early adopter of evaluation and performance measurement. Beginning as early as the 1970s, USAID initiated use of both qualitative and quantitative program evaluation to demonstrate project impact. The USAID example

BOX 3
Foreign Aid Example of The Micro-Macro Paradox

For a quarter of a decade, economists have debated the effectiveness of foreign aid. The term "micro-macro paradox" was introduced by Paul Mosley (1986) to explain the gap between the micro-level impact of individual development projects, often measured through program evaluation and monitoring and evaluation and the macro-level impact. Econometrics typically utilizes large-N methods with regression analyses to determine macro-level outcomes.

The debate continues. For example, H. White (1992) found that of 245 projects evaluated by the World Bank, 85 percent of the World Bank's projects did well, but that there was no measurable change at the macroeconomic level. More recent comparisons of World Bank and Asian Development Bank projects finds that half to over four-fifths of projects were successful, while there continues to be questions about macro-economic impact (M.G. Quibria (2004)).

Some more recent research finds a modest macro-economic impact, while others argue that there are other factors and a broadening array of development goals (e.g., poverty eradication). This debate demonstrates the challenges of linking micro and individual level changes to macro-outcomes.

M.G. Quibria "Development Effectiveness : What Does Recent Research Tell Us?" 2004. Asian Development Bank. Working Paper #1.
http://www.adb.org/documents/OED/Working-Papers/oct01-oed-working-paper.pdf
H. White, H. 1992. "The Macroeconomic Analysis of Aid Impact." *Journal of Development Studies* 28.
Paul Mosley "Aid-effectiveness: The Micro-Macro Paradox" IDS Bulletin 17 (No. 2). Pp. 22 - 27
Institute of Development Studies

demonstrates how much federal agencies aspire to making macro-level conclusions. For example, the goal is not to say how political leaders were trained on legislative strengthening or on the use of gender quotas or their extent to which they found the training helpful, but rather, to what extent countries receiving aid have become more democratic.

USAID has created a detailed database of USAID Democratic Governance (DG) expenditures made since 1990 and sponsored a study of the impact of USAID democracy assistance which found that its programs have had a measurable, positive impact on democratic progress around the world (Finkel *et al.*, 2007; 2008). These researches found that for every additional $10 million dollars invested in democracy assistance, the recipient country is predicted to gain one quarter of a point on Freedom House democracy index. This study utilizes an econometric-style analysis of macro-level impact, and not performance measurement methods.

However, this method of determining macro-level impacts is not entirely persuasive. According to a 31-member panel of experts reviewing USAID and other democracy providers, "Despite many billions of dollars and the sustained attention of thousands of PhD's in the World Bank, academia and elsewhere, it is not clear that foreign assistance has done any good at all....If foreign assistance, including democracy assistance, was a company, it would have gone into bankruptcy twenty years ago" (CDDRL, 2008:10).

The promise of performance measurement is a powerful one – the ability to draw inferences about agency-wide or accountable, geographic results provides the strongest justification for instituting new ways to measure and evaluate government programs. The problem is that the gap between knowing we need to adopt the common sense solution – and actually devising and implementing an evidence-based program or policy – is large and growing. It is this gap that this book seeks to address.

Performance measurement is a research method in its infancy. Performance measurement may never fully link micro- to macro-level outcomes. But the promise still offers new ways to measure effective programs and policies and its reach remains underutilized.

Promise #2: Providing for Democratic Accountability

There are many definitions of accountability – including accountability "to" and accountability "for" criteria. I terms these "complex" definitions which go beyond the basic content. These focus on

1. To "whom" accountability is due (e.g., President, Congress, citizens);
2. The "mechanism" of acccountability (e.g., reporting, standing for election);
3. The "arena" of accountability (e.g., fiscal, legal, political, performance).

Complex definitions are diverse and make the most sense within a specific context – e.g., how members of Congress provide political accountability to their constituents by campaigning and being elected, or how public administrators provide legal accountability to Congress by reporting how they follow congressional mandates.

At its base, however, accountability means simply whether the measures are able to provide a factual accounting of major activities that are its focus. In this narrow sense, accountability rests upon the outcome measure's ability to truthfully account for desirable key features of the activities. In this regard, performance evaluation promises to provide an objective basis for measuring the outcomes of public policies.

This is a promise that is shared with other research methods – but it is one that is provided using different research methods, and this is the argument to which I turn next.

The Research Methods Justification of Performance Evaluation

The basic sine qua non of a research method is the ability to draw inferences. If attributions and inferences are the "Achilles' heel" of performance measurement (Hendricks, 2000), it is performance evaluation as a research method that should be able to answer this existing weakness. Michael Hendricks has aptly noted, attribution is the "Achilles' heel" of performance measurement.

Causal reasoning is of the form that "X" causes "Y." The various types of established causal reasoning used in the research process include neo-Humean regularity theory, counterfactual theory, manipulation theory, and mechanisms and capacities theory (Brady, 2002). In this paper, I argue that linking together steps in a process also provides a form of causal linkage. Managing for performance is based upon the assumption that there is a process that – if managed well – and implemented with fidelity – can produce the outcomes better and more efficiently. In jumping to measurement of outcomes, managers are unwittingly skipping attention to the methods. And performance outcome measures are more than a social indicator – they are a specific type of social indicator of populations that has a causal linkage with the program's activities. Each of these types of linkages require methodologically grounded inferences to establish validity and reliability. When the business process approach is transferred to the public sphere without a methodological understanding, mistakes can occur. This is particularly important in a field where cross-disciplinary methods (experimental, survey, interview, focus groups, econometric) and data sources (original as well as documentary data collection) must be integrated.

I will here only outline the basics of this argument. It includes four measurement models, three inferences, and four analytical issues that must be considered in performance evaluation. The first is a cultural and institutional distinction, while the inferences and analytical issues are research methods questions.

Institutions, Languages and Measurement Models

It is important to understand that there are four measurement models bound up with performance measurement: policy, program, management, and accountability strategies. Each of these four measurement models have different languages and different approaches to measurement. Critical to understanding these measurement models is the fact that each are measuring different aspects of the same activity and the activities when linked as a process over time. As such, each presumes difference inferences about activities at the policy, program, management, and accountability levels. Performance

measurement is designed to support the policy process *as it is being made*. Therefore, anyone wishing to master performance measurement must understand the policy process *at the leadership levels where policy is made, using models that leaders themselves use to define and make policy*. The problem is that there is not just "one" policy process. At the leadership level, there are four basic policy cultures – each of which speak very different policy languages and manage the policy process from very different vantage points using distinct processes

Business and the language of programs. The language most unfamiliar to those working in policy communities is that spoken in the business world. Business organizations speak the language of programs. Actually, the term for programs in the private business world is "business process." The business process are the steps for creating a product of value. Here, strategy is driven by the marketplace, which has its own dynamic. Without a monopoly, a free market cannot be controlled by individual business enterprises. Instead, a competitive business must need the self-defined needs of the marketplace in order for this product to have value. The parallel for government in the public sphere is the more neutral "programs." In programs, the goal is to ensure a result despite the "noise" of other factors. Both arenas of action – public and private – seek (or should seek) to manage the activities or steps in a program or a business process. It is the world of business, however, that has actually developed management standards for effectively managing the business process. It is not so often recognized that major standards in public management are derived from those innovated first in business. While governments do not seek to earn profits, they do run programs that are designed to be effective with specific individuals or communities just as business organizations seek to gain a competitive advantage by producing a product valued by the market. Understanding the language of the business process is essential for performance measurement. This is a language where the focus is local, the context specific, and the goal is to ensure that all steps in a program are designed to meet a concrete outcome.

Policy makers and the language of policy. Policy makers speak the language of policy. This is the language that will be most familiar to those with political science and public administration backgrounds since political scientists teach and write about political strategies. Policy makers respond to democratic political conflict and constituencies. Their world centers around setting goals focused on solving concrete problems. The language of policy is spoken by leaders whose mandate is earned through elections and political appointments (whether administrative or judicial). The language of policy is typically spoken by politicians and elected officials within legislatures who are developing new policies to address problems. Here, "policy" itself is understood as a "strategy;" i.e., a newly formulated policy reflects a new approach to solve a problem. For example, the COPS (Community Oriented Policing) program was designed to place officers in neighborhoods who could then address local neighborhood problems. This is a language based on logrolling, coalition-building and representation (e.g., constituency mobilization, campaigns, town hall meetings) and legislative strategies (e.g. committee hearings, oversight, investigation, creation of new authorities, budgeting) centered on current public policy problems. This language, common in political arenas such as campaigns and legislatures, stresses goals and outcomes.

Among policy makers, there are be a variety of different subcultures. Clearly, Democrats and Republicans have very different, polarized cultures with opposing constituencies focused on different aspects of social problems. To give another example, as many observers have noted, the U.S. House and the U.S. Senate have very different cultures. The saying is that in the House, the individual member conforms to the body,

while in the Senate, the body conforms to the individual member. This saying encasulates a number of institiutional and procedural differences between the two houses. Yet, Senators and Representatives and Democrats and Republicans all speak the language of policy, as do mayors, governors, state legislators, and judges.

The academy and the language of accountability. The language of accountability is spoken in universities – the "academy." The reason why this language is termed "accountability" is because this is the one language most concerned with what provides the most accurate information and knowledge needed for accounting (the basic definition). This language is based on the principles of free inquiry and the scientific method. The ability to acquire the language of accountability requires considerable advanced graduate training. There are an increasing number of professional evaluators who possess this advanced training. As discussed earlier, the driver of the academy is the scientific method. Only the "gold standard" of the scientific method can provide good (or better) knowledge. The "gold standard" is the ideal of experimentation, because it alone provides the strongest evidence of whether X causes Y.

Public administration and the language of management. Finally, the language of public management is spoken by public administrators. In public administration, traditional accountability are administrative including "*hierarchical accountability for inputs* (administrative rules guiding routine tasks and budgetary allocations) and [legal] *legal accountability for processes* (audits, site visits, and other monitoring tasks)" (Heinrich, 2002: 721-22; emphasis in the original). Public administrators are civil servants who work in public agencies at federal, state and local levels. What Wamsley and his colleagues call "The Public Administration," is "self consciously derived from, and focused upon... an Agency Perspective" (P. 36). This "Agency Perspective" goes beyond experience in managing in a political context to include competencies in governance based upon "an historic, covenantal, organic and constitutional" perspective (p. 39). The expertise of "The Public Administration" reflects a more long-term perspective on "what works" – beyond the sort-term measures of performance, a non-political neutrality on implementing official public policy. Here, the driver is neutral competence within the rule of law, not the political process. As such, The Public Administration is designed to serve within official policies and decisions made by democratically elected officials.

Research Methods of Performance Evaluation

There are several technical issues related to defining the research methods of performance evaluation. The details are provided in the Appendix, and discussed below.

Three Inferences. There are three types of *inferences* that a valid, reliable and responsive performance measurement system must encapsulate:

- About **relationships** – *how are two outcomes linked?*
- About **outcomes** – *how good are the indicators of outcomes?*
- About a **process** – *what are the events and activities that produce outcomes?*

Each of these reflects different assumptions and definitions of "truth" and "success" which are best understood through six analytical issues.

Issue #1. Validity and Reliability are Different. Validity and reliability are different, but linked ideas. While highly developed concepts in scientific thinking, they are commonsense ideas as well. In science, they are *different* because they ask different questions:

- The key validity questions asks whether the desired phenomenon is actually being measured. For example, does the Stanford-Binet IQ test measure "intelligence" – or are there multiple types of that concept, e.g., artistic skills as opposed to cognitive skills?

- The key reliability question asks whether the measurement of the same phenomenon consistent or repeatable over time? For example, if you measure temperature repeatedly over a 24 hour period, are the changes you find due to variations the thermometer or to a fever?

The concepts are *linked* because to be a valid measure, it must also be reliable. However, an instrument can be reliable (i.e., give consistent measures) – but not a valid measure of a concept. Validity and reliability are also *linked* because – from a scientific perspective – both are considered to be *inferred* concepts. This means that they are never proven or known directly. Thus, in scientific instrumentation, reliability is never fully known – it also is a "relative" concept compared to an unknown or ideal "true" score. The degree of reliability is identified when you can infer "low error" against this ideal. Reliable findings assume that the "observed score" or "observed relationship" consists of three components:

"True" score + Systematic error + Random error

And in scientific theory, validity is never "proven" because it is inferred from evidence from the research and sampling designs through a process of consensus by experts in the field who are familiar with a variety of research. Valid results are assumed to reflect universal relationships and genuine outcomes.

Issue # 2. Science and Business Process Management Define These Terms Differently. Validity and reliability ideas are common to both science and business process management. Both define these ideas quite differently – and yet both methods are needed for performance evaluation. Science has a metaphysical notion of "truth" as something that can never be fully known. While individual scientists compete for making scientific discoveries, they work within an agreed upon body of scientific knowledge that is developed consensually with colleagues. In contrast, the business model focuses on "what works" – based on expedience and taking advantage of opportunities against competitors to develop unique "market share." Business uses a verification model. Instead of using an elaborate method, business relies on management of all details that seeks to manage a very well-defined process. Science uses an elaborate model of testing for validity and reliability based on what is known as the "logico-empirical" model. This model produces an entirely different type of "explanation" that includes just a few factors that provide a parsimonious explanation of universal relationships – relationships among concepts that exist across time and space. To give an example, the business model would try to produce a list of the hundreds of factors that produced the French Revolution, while science would look for a few factors that all revolutions have in common. The business model seeks an

ideographic explanation, while the science model seeks a parsimonious explanation. Both are valuable – but very different approaches to "truth" – and different approaches to validity and reliability. And both are needed in performance measurement which focuses on producing a distinct outcome in a specific situation.

Issue # 3. Performance Evaluation is an Applied Model Drawing Upon Both Science and Business Models. Scientific models examine both relationships and develops indicators for outcomes. These are causal determinations. Is the relationship being measured and do the indicators faithfully reflect the outcomes? The business process is a different thing conceptually and from a research perspective. It requires different tools for identifying and managing, including the identification of a repeatable process for assessment, management and reengineering. Key to understanding processes is that they are different from a "causal" factor – they might be better understood as developmental in nature. Each analytical issue is detailed in a separate chart that describes the similarities and differences of each set of assumptions. Social science provides for different types of "scientific" explanation: the business process method requires the "genetic" model of explanation in which each outcome is affected by the prior outcome.

Issue # 4: As a Result, There are Three Types of Inferences that Need to be Drawn for Performance Measurement. The three types of inferences are relationship, outcomes/indicators and the business process. The theory of change and the matrix of events and measures provide tools for linking relationships. A separate inference, the logic model, provides the basis for the inference of short-term, intermediate and long-term outcomes. Finally, the inference of the business process is demonstrated by the results chain and indicates how processes are linked to outcomes.

Issue #5: What You Need to Demonstrate Varies According to the Type of Inference You Wish to Draw. Each inference – validity and reliability – is based up a different demonstration. Valid relationships are demonstrated through the counterfactual, while reliable relationships are demonstrated through universal relationships. Valid outcomes and indicators are demonstrated through generalizability, while reliable indicators are demonstrated through consistency. For the business process,, a valid process is demonstrated through either identification of a core process or link to a strategic (goal-centered) process. A reliable business process is one that is defined rather than an ad hoc one.

Issue # 6: Methods and Criteria Used to Demonstrate Validity and Reliability Are Well-Defined According to the Type of Data Inference Being Justified. And each type of data inference is, in turn, linked with specific methods that are related to each type of relationship that must be demonstrated. This includes measuring inputs, outputs and outcomes for personnel, programs, policies and strategies, and can extend into the AS-IS and TO-BE models. For programs, the two comparisons should be identical (program implemented with fidelity). The analysis would differ for policies and strategies. Ideally, both need to be managed to ensure achievement of results.

TABLE 3
PERFORMANCE EVALUATION OF PROGRAMS, POLICIES AND STRATEGIES: TYPES OF INPUTS, OUTPUTS AND OUTCOMES TO MEASURE

	INPUTS Resources	**OUTPUTS** Result of Action	**OUTCOMES** Normative Goal
PERSONNEL *Staff Working Within a Program or Unit on Similar Objectives*	Workload Hours Worked	Counts of Work Products Related to Goal or Program Outcome	Achievement of Work Product Benchmarks; "Customer" Satisfaction; Supervisor Ratings of Quality; Relative Success Compared to Colleagues
PROGRAM *Discrete Entity with devoted funding*	$$ Personnel	Counts of Defined Program Activities	Progress Toward Program Goal Achieved (e.g., change in social outcomes)
POLICY *Official and/or Formal Decisions that codify strategies*	Laws Regulations Official Statements	Counts and Types of Implementation Activities	Change in Priorities Invested, Embedded and/or Achieved (e.g., new constituencies, new resources, new and/or better programs, etc.)
STRATEGY *Position and Perspective on Policies that orient management or executive action*	Plan (prospective) or Pattern (retrospective)	Counts of New Capacities, Infrastructures, Programs, Program Activities or Policies Created	Government or Program Reform Achieved (more effective or efficient government or new cross-cutting goal achieved)

Conclusions

In this paper I have argued that we need to recognize and developed the distinctive research methods of performance evaluation – separate from program evaluation and separate from existing "scholarly" mono-methods. We currently lack a defined model, but one can be created by recognizing the importance of how validity is different from reliability and that there are not only trade-offs between the two, but they are defined differently in "science" and according to the "business process" which is the "new kid" on the evaluation block.

What is being missed by this lack of a defined model? The simple answer is that we are

probably not getting the results that our elected officials are promising us. The more complex answer to the what is missed question is that research methodologists have missed defining evaluation as anything more than a mechanical process where any measure will do. In fact, measurement validity (Adcock and Collier, 2001) is distinct from causal validity for which quantitative scholars and philosophers of science have identified at least four types (Brady, 2002). In failing to elucidate the methods and distinguishing the research process for performance evaluation from the management and the policy roles, methodologists have ignored the fact that the "business process" introduced as a central part of the performance orientation to policymakers in the 1990s is a new form of causal reasoning used in performance evaluation.

And why does it matter? The simple answer is that it matters because the accountability provided by objective science provides the best information. This is critical to ensure neutral competence plays a role in the provision of public goods alongside what E.E. Schattschneider called the democratic collaboration of "ignorant people and experts". The nature of objective research is important and needs to be understood within its own language and standards before comparing it to other ways of valuing and making policy choices such as political accountability or budgetary or management considerations.

The more complex answer to the why does it matter question is that if we do not understand the method of performance evaluation, then we are at risk of losing meaningful political accountability. Political and democratic accountability comes through elected officials and the realignment of political agendas when electoral mandates are either changed or reinstated through elections. As B. Guy Peters points out, "If civil servants and other appointed officials are indeed to become entrepreneurial then they must become less dominated by the dictates of these [political] masters. If this approach were practiced, it would alter fundamentally ideas of [democratic] accountability" (2001: 8). Not only are federal managers increasingly empowered to make agenda choices and to manage with fewer restrictions with the new 2010 GPRAMA which asked managers to define and focus on a small set of agency priorities rather than trying to work equally across all programs within their jurisdiction. And to the degree that what civil servants (and outside consultants) make these political as well as evidence-based decisions in ways that lack basic accountability about objective results (what works) as well as an increasing freedom from political accountability to elected officials (and voters), democracy is seriously threatened.

References

Almond, Gabriel A. 1988. " Separate Tables: Schools and Sects in Political Science" PS: Political Science and Politics 21: 828-842

Alkin, Marvin C. and Christina A. Christie. Alkin, M. C. (2012). Evaluation roots (2nd ed.). Thousand Oaks, CA

Bachrach, Peter and Morton S. Baratz. 1963. Decisions and NonDecisions: An Analytic Framework. *American Political Science Review* 57: 632-42.

Baer, Denise. Forthcoming, 2009. *Delivering Measurable Performance: Methods, Strategies and tools for Policymaking and Public Management*. Thousand Oaks, CA: SAGE Press.

Barley, Zoe A. and Mark Jenness. 1995. *Conceptual Underpinnings For Program Evaluations Of Major Public Importance: Collaborative Stakeholder Involvement*. In Joy A. Frechtling (Ed). *Footprints: Strategies for Non-Traditional Program Evaluation*. Washington, DC: Division of Research, Evaluation and Dissemination. National Science Foundation. Pp. 97-106.

Baumgartner, Frank D. and Bryan D. Jones. 1993. *Agendas and Instability in American Politics*. University of Chicago Press.

Brewer, Garry D. 1974. "The Policy Sciences Emerge: To Nurture and Structure a Discipline." *Policy Sciences* 5: 239–244.

Burstein, Paul. 1991. "Policy Domains: Organization, Culture, and Policy Outcomes." In W. Richard Scott and Judith Blake (eds.), *Annual Review of Sociology*. Volume 17. Palo Alto, Calif.: Annual Reviews, Inc.

Beetham, David. 2004. "Towards a Universal Framework for Democracy Assessment." *Democratization* 11: 1-17.

Bollen, Kenneth, Pamela Paxton and Rumi Morishima. 2005. Research Design to Evaluation the Impact of USAID Democracy and Governance Programs." Typescript accessed January 24, 2009 at: http://hedprogram.org/Portals/0/RFA%20docs/SSRC%20research_design_final_version.pdf

Bourdeaux, Carolyn and Grace Chikoto. 2008. "Legislative Influences on Performance Management Reform." Public Administration Review 6: 253-265

Boyne, George A., Kenneth J. Meier, Laurence J. O'Toole, Jr., and Richard M. Walker. 2005. "Where Next? Research Directions on Performance in Public Organizations." Journal of Public Administration Research and Theory Vol. 15: 633-639

Brady, Henry and David Collier (Eds.). 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman and Littlefield.

Brady, Henry E. and Jason Seawright. 2004. "Framing Social Inquiry: From Models of Causation to Statistically Based Causal Inference" A Paper prepared for the American Political Science Association Annual Meetings, Chicago, Illinois.

Bryman, Alan. 2006. "Integrating Quantitative and Qualitative Research: How is it Done?" *Qualitative Research* 6:97-113.

Burnell, Peter (Ed). 2007. *Evaluating Democracy Support: Methods and Experiences*. International Institute for Democracy and Electoral Assistance and Swedish International Development Cooperation Agency

Bjuremalm, Helen. 2006. "Power Analysis: Experiences and Challenges. A Concept Note." SIDA. Available Online at: http://www.odi.org.uk/rapid/Tools/Toolkits/Mapping_Political_Context/Power_analysis.html

Cartwright, Nancy and Jeremy Hardie. 2012. Evidence-Based Policy: A Practical Guide to Doing It Better. New York: Oxford.

Clapp–Wincek, Cynthia and Richard Blue. 2001. "Evaluation of Recent USAID Evaluation

Experience." Submitted to U.S. Agency for International Development. PPC/CDIE

Collier, David, Henry E. Brady, and Jason Seawright, "Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology,"

Clarke, Kevin A. 2007. "The Necessity of Being Comparative:Theory Confirmation in Quantitative Political Science." *Comparative Political Studies* 40: 886.

Clarke, Kevin A. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22.

Crenson, Mathew A. 1971. *The Un-Politics of Air Pollution: A Study of Non-Decisionmaking in the Cities .* Johns Hopkins University Press.

Cresswell, J. and V.L. Plano Clark. 2007. *Conducting and Designing Mixed Methods Research*. Thousand Oaks, CA: Sage.

Dehue, Trudy. 2001. "Establishing the Experimental Society: The Historical Origin of Social Experimentation According to the Randomized Controlled Design." American Journal of Psychology 114: 282-302.

Dery, David. 1984. "What is a Problem, so that it May be Usefully Defined?" and "Social Problems as Opportunities for Improvement." Pp. 21–36 in *Problem Definition in Policy Analysis*. Lawrence: University Press of Kansas.

DFID. 2008. *Using Drivers of Change to Improve Aid Effectiveness*. United Kingdom Department for International Development. Available Online at: http://www.dfid.gov.uk/aboutDFID/organisation/driversofchange.asp

Dubnick, Melvin J and H. George Frederickson. 2011. Accountable Governance: Problems and Promises. New York: M.E. Sharpe.

Faure, Murray Andrew. 1994. Some Methodological Problems in Comparative Politics. *Journal of Theoretical* Politics 6: 307-322

Frechtling, Joy A. (Ed.) 1995. *Footprints: Strategies for NonTraditional Program Evaluation*. National Science Foundation. Pp. 25-36.

Frechtling,. Joy A. (Ed.). 1995. *Footprints: Strategies for Non-Traditional Program Evaluation*. Washington, DC: Division of Research, Evaluation and Dissemination. National Science Foundation.

Frederickson, David G. And H. George Frederickson. 2007. Measuring the Performance of the Hollow State. Washington, D.C.: Georgetown University Press.

Gerber, Alan and Donald Green. 2002. .Reclaiming the Experimental Tradition in Political Science, In *Political Science: State of the Discipline*, edited by Ira Katznelson and Helen V. Milner, New York: W.W. Norton, pp. 805-832

Gerring, John. 2004. .What is a Case Study and What is it Good for?. *American Political Science Review* 98: 341-354.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." Political Research Quarterly 52: 647-74

Gorman, D.M., and Erich W. Labouvie Using Social Indicators to Inform Community Drug and Alcohol Prevention Policy."Journal of Public Health Policy 21: 428-446.

Green, Donald P. and Alan S. Gerber. 2008. *Get Out the Vote, Second Edition How to Increase Voter Turnout.* 2nd Ed. Brookings Institution.

Heineman, Robert A., William T. Bluhm, Steven A. Peterson and Edward N. Kearny. 1990. The World of the Policy Analyst: Rationality, Values & Politics. Chatham, NJ: Chatham House.

Heinrich, Carolyn J. And Laurence E. Lynn, Jr. 2000. Governance and Performance: New Perspectives. Washington, D.C.: Georgetown University Press.

Howlett, Michael and M. Ramesh. 2003. *Studying Public Policy: Policy Cycles and Policy Subsystems* 2nd ed.. Toronto: Oxford University Press.

Christopher Hood, Christopher and Ruth Dixon. 2010. The Political Payoff from Performance Target Systems: No-Brainer or No-Gainer? Journal of Public Administration Research and Theory 20: 281-298

Judicial Watch. 2011. "President Obama's Czars: A Judicial Watch Special Report." September 15, 2011. On the web at: http://www.judicialwatch.org/files/documents/2011/czar-report-09152011.pdf

Kanter, Rosabeth Moss and Derick Brinkerhoff. 1981. "Organizational Performance: Recent Developments in Measurement." Annual Review of Sociology 7: 321-349

Kellogg, WK Development Foundation. 2004. Logic Model Development Guide: Using Logic Models to Bring Together Planning, Evaluation and Action. Battlecreek, MI: WK Kellogg Foundation. Available online at: http://www.wkkf.org/DesktopModules/WKF.00_DmaSupport/ViewDoc.aspx?LanguageID=0&CID=281&ListID=28&ItemID=2813669&fld=PDFFile

Kettinger, William J., James T. C. Teng, and Subashish Guha. 1997. "Business Process Change: A Study of Methodologies, Techniques, and Tools." MIS Quarterly 21: 55-80

King. Gary. 1998. Unifying Political Methodology: The Likelihood Theory of Statistical Inference. Ann Arbor: University of Michigan Press.

Klitgaard, Robert and Paul C. Light. 2005. High Performance Government: Structure, Leadership, Incentives. Santa Monica, CA: Rand Pardee Rand Graduate School

Kusek, Jody Zall and Ray C. Rist. 2004. *A Handbook for Development Practitioners: Ten Steps to a Results-Based Monitoring and Evaluation System*. Washington, DC: World Bank.

Langbein, Laura with Claire L.Felbinger. 2006. *Public Program Evaluation: A Statistical Guide*. Armonk, NY: M. E. Sharp

Chalmer E. Labig, Chalmer #. 2009. "Bad Measures Don't Make Good Medicine: The Ethical Implications of Unreliable and Invalid Physician Performance Measures." Journal of Business Ethics, Vol. 88: 287-295.

Light, Paul C. 2005*. The Four Pillars of High Performance: How Robust Organizations Achieve Extraordinary Results. Lessons From the Rand Corporation*. New York: McGraw Hill.

Light, Paul C. 2005. The Four Pillars of High Performance: How Robust Organizations Achieve Extraordinary Results. New York: McGraw Hill.

McGlynn, Elizabeth A. 2003. An Evidence-Based National Quality Measurement and Reporting System." Medical Care 41: I8-I15

Mansfield, Edward and Snyder, Jack. 2007. "The Sequencing Fallacy." *Journal of Democracy* 18: 5–10

McFaul, Michael Amichai Magen & Kathryn Stoner-Weiss. 2008. "Evaluating International Influences on Democratic Transitions: Concept Paper." Center on Demcracy, Development and the Rule of Law, Stanford University http://iis-db.stanford.edu/res/2278/Evaluating_International_Influences_-_Transitions_-_Concept_Paper.pdf

McLaughlin, John A. and Gretchen P. Jordan. 2004. "Using Logic Models." In Joseph S. Wholey, Harry P. Hatry, and Kathryn Newcomer (Eds.) *Handbook of Practical Program Evaluation*. 2nd Ed. Jossey Bass. pp. 7-32.

Millett, Ricardo A. 1996. "Empowerment Evaluation and the W.K. Kellogg Foundation,"

Empowerment Evaluation." In *Knowledge and Tools for Self-Assessment & Accountability*, David Fetterman, Shakeh Kaftarian, and Abraham Wandersman (eds.), Sage Publications, Inc., Pp. 65-76.

Morton, Rebecca and Kenneth Williams, 2010. Experimental Political Science and the Study of Causality: From Nature to the Lab. New York: Cambridge.

Moynihan, Donald P. And Patricia W. Ingraham. 2003. "Look for the Silver Lining: When Performance-Based Accountability Systems Work." Journal of Public Administration Research and Theory 13:469-490

Donald P. Moynihan, Sanjay K. Pandey. 2010. "The Big Question for Performance Management: Why Do Managers Use Performance Information?" Journal of Public Administration Research and Theory 20: 849-866

National Academies Press. 2008. *Improving Democracy Assistance: Buildign Knowledge through Evaluations and Research*. Washington, D.C.: National Academies Press.

Newcomer, Kathryn E. 2004. "How Might We Strengthen Evaluation Capacity to Manage Evaluation Contracts?" *American Journal of Evaluation* 25: pp. 209-218.

Newcomer, Kathryn E. And Mary Ann Scheirer. 2001. "Using Evaluation to Support Performance Management: A Guide for Federal Executives." Washington, DC: IBM Center for the Business of Government

Quibria, M. G. 2004. *Development Effectiveness: What Does Recent Research Tell Us?* Asian Development Bank.

Robinson, W.S. 1950. "Ecological Correlations and the Behavior of Individuals". *American Sociological Review* 15: 351–357.

Romano, Patrick S. and Ryan Mutter. 2004. "The Evolving Science of Quality Measurement for Hospitals: Implications for Studies of Competition and Consolidation." International Journal of Health Care Finance and Economics 4; 131-157.

Rossi, Peter. 1987. "The Iron Law Of Evaluation And Other Metallic Rules." Research in Social Problems and Public Policy 4: 3-20.

Rossi, Peter and Howard Freeman. 1985. Evaluation: A Systematic Approach (3rd ed.). Beverly Hills, CA: Sage.

Russ, Darlene-Eft, Marcie Bober, Ileana de la Teja, Marguerite J. Foxon, and Tiffany A.Koszalka. 2008. Ealuator Competencies: Standards for the Practice of Evaluation in Organizations. San Francisco, CA: Jossey Bass.

Sabatier, Paul A. (Ed). 2007. 2nd Ed. Theories of the Policy Process. Boulder, CO: Westview.

Sarles, Margaret J. 2007. " Evaluating The Impact and Effectiveness of USAID's Democracy and Government Programmes." In Peter Burnell (Ed.) *Evaluating Democracy Support: Methods and Experiences.* SIDA and International IDEA. Pp. 48-70.

Schwartz-Shea, Peregrine and Dvora Yanow. 2002. "Reading" "Methods" "Texts": How Research Methods Texts Construct Political Science." Political Research Quarterly 55: 457-486.

Sherman, L., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., and Bushway, S. 1998. *Preventing Crime: What Works, What Doesn't, What's Promising – A Report to the United States Congress*. Prepared for the National Institute of Justice.

Smith, Keving B. And Christopher W. Larimer. 2009. The Public Policy Theory Primer. Westview Press.

Smith, Rogers M. 2002. "Should We Make Political Science More of a Science or More About Politics?" *PS: Political Science and Politics* 35:199-201.

Stinchcombe, Arthur. 1968. *Constructing Social Theories.* Chicago: University of Chicago Press.

Stokey, Edity and Richard Zeckhauser. 1978. A Primer for policy Analysis. New York: W.W. Norton.

Stoto Michael A. 1997. "Methodological Issues in Developing Community Health Profiles and Performance Indicator Sets." In Durch, Jane S., Linda A. Bailey and Michael A. Stoto (Eds.). 1997. *Improving Health in the Community: A Role for Performance Monitoring*. Washington, D.C.: National Academy Press.

Thompson, James. R. 2000. "The Dual Potentialities of Performance Measurement: The Case of the Social Security Administration." *Public Productivity & Management Review* 23: 267-281.

Tassakori, A. And C. Teddlie. 2003. *Handbook of Mixed Methods in Social and Behavioral Research*. Thousand Oaks, CA: Sage.

USAID (United States Agency for International Development). 1998. Handbook of Democracy and Governance Programme Indicators. Available on the Web at: http://www.usaid.gov/our_work/democracy_and_governance/publications/pdfs/pnacc390.pdf].

Wantchekon, Leonard. 2003. .Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin,.World Politics 55: 399-422.

Weimer, David L. And Aidan R. Vining. 2011. Policy Analysis. Boston: Longman.

Weiss, Heather B. 2001. "Reinventing Evaluation to Build High-Performance Child and Family Interventions." In *Perspectives on Crime and Justice: 1999-2000 Lecture Series*. NCJ 184245. National Institute of Justice.

Wholey, Joseph S. 1979. Evaluation: Promise and Performance. Washington, D.C.: The Urban Institute.

World Bank. 2004. Monitoring *& Evaluation: Some Tools, Methods and Approaches*. Washington, DC: The World Bank.

Yin, Robert K. 1995. "New Methods For Evaluating Programs In NSF'S Division Of Research, Evaluation And Dissemination."

APPENDIX

KEY ANALYTICAL ISSUE # 1
Validity and Reliability are Different

	Validity	Reliability
Different Questions are Addressed	Validity Question: Is the desired phenomenon actually being measured? Example: Does the Stanford-Binet IQ Test measure "intelligence" or are there multiple types?	Reliability Question: Is the measurement of the same phenomenon consistent or repeatable over time? Example: Is the variation in temperature due to the thermometer or to a fever?
How The Concepts are Linked	To be valid, a measure MUST be reliable	A reliable instrument may NOT be valid
Both are Inferred Concepts	Validity is never "proven" – it is inferred from evidence from the research and sampling designs through a process of consensus.	Reliability is never fully known – it is a relative concept. The degree of reliability is identified when you can infer "low error."
Different Assumptions are Made	Valid findings assume that the finding is measured appropriately when there is a consensus over the results. Valid results are assumed to reflect universal relationships and genuine outcomes.	Reliable findings assume that the "observed score" or "observed relationship" consists of three components: "True" score + Systematic error + Random error Reliable results have "low" error.
Different Types of Evidence are Required to Draw an Inference	***Validity is inferred when you can demonstrate:*** Internal Validity: Does X "cause" Y? (Demonstrated by how well the Research Design addresses identified threats to validity) External Validity: To what types of settings, populations and variables can the findings be generalized to? (Demonstrated by how well the Sampling Design allows you to generalize): Face Validity – consensus among laypersons, non-experts, citizens, general public Content Validity – consensus among experts in the field Predictive Validity – accurately predicts future behavior Concurrent Validity – results consistency with other indicators seeking to measure the same concept Construct Validity – consistency with theories	***Reliability is inferred when you can demonstrate:*** Equivalence – agreement between two or more instruments administered at nearly the same time *(e.g., the Parallel or Alternative Forms Test, the Inter-Rater Test)* Stability – same results obtained with repeated testing of the same design or instrument on the same sample or population (e.g., Test-Re-Test Reliability) Internal consistency or Homogeneity – various items in a single measure administered to the same people at the same time all appear to reflect the same attribute (i.e., give similar results).

KEY ANALYTICAL ISSUE # 2

Science and Business Process Management Define These Terms Differently

	Factor		Scientific Meaning	Business Process Meaning
Validity	Relationship		Universal relationship regardless of setting	Defined managed relationship dependent on setting and distinct characteristics of management and staff
	Type of Explanation		Parsimonious	Idiographic
	Criteria for Drawing Conclusions		Cautious criteria for establishing facts – may take years to draw conclusions	Expedient criteria for making decisions – need to respond to events even in the face of limited information
	Definition of "Truth"		Never known	Proven by events
	Basis for Stating "Truth"		Metaphysical Use of logico-empirical paradigm for drawing inferences	Expedient Opportunistic Entrepreneurial
	Posture	Present	Backward Looking	Use current opportunities to create new future
		Peers	Colleagues – participate to develop a common body of scientific knowledge following a common set of rules using peer review.	Competitors – compete by taking advantage of opportunities and weaknesses among competitors by developing market share.
Reliability	Equivalence of forms or observers		Equivalence	Management Defined Process Run - ReRun Statistical Process Control
	Stability of results		Test-ReTest stability	
	Homogeneity of individual elements		Split Halves/Cronbach's Alpha	

KEY ANALYTICAL ISSUE #3:

Performance Evaluation is an Applied Model Drawing Upon Both Science and Business Models

MODEL		EXAMPLES	MAJOR SOURCES	BENEFITS	COMMENT
Source	Type				
Three Fundamental or Systematic Models					
SCIENCE	Relationships	***Blue Prints:*** Model Violence Prevention Programs *On the Web at:* http://www.colorado.edu/cspv/blueprints	Any research methods textbook	Provides valid and reliable information about "what works."	Requires demonstration of causality based on principles of internal and external validity
	Outcomes and Indicators	***Kids Count*** Annie E. Casey Foundation Data Book *On the Web at*: http://www.aecf.org/kidscount/	Any research methods textbook	Identifies good indicators measuring outcomes.	Requires methods to establish reliability and validity
BUSINESS	Process	Often evident when new models are introduced: e.g., traditional policing vs. community or problem-oriented policing	William J. Kettinger, James T. C. Teng, Subashish Guha. 1997. "Business Process Change." *MIS Quarterly* 21: 55-80	Tools for identifying business processes.	Requires identification of a repeatable process for assessment, management and reengineering
Two Applied Models					
HYBRID –	Program Evaluation	Evaluations of the DARE Program "Youth Illicit Drug Use Prevention: DARE Long-Term Evaluations and Federal Efforts to Identify Effective Programs, GAO-03-172R" *On the Web at*: http://www.csdp.org/news/news/darerevised.htm	Peter H. Rossi, Mark W. Lipsey and Howard E. Freeman. 2003. *Evaluation: A Systematic Approach*. Sage Publications	Identifies which existing programs work and which do not based upon scientific criteria.	Requires an "evaluable" program with a consistent and reasonably well-understood process or program.
HYBRID –	Performance Evaluation	Six Sigma *On the Web at:* http://www.isixsigma.com/ ISO 9001 *On the Web at:* http://www.iso.ch/iso/en/ISOOnline.frontpage Baldridge Criteria for Performance Excellence *On the Web at:* http://www.quality.nist.gov/	Hatry, Harry. 1999. *Performance Measurement: Getting Results*. Washington, D.C.: Urban Institute. United Way. *Measuring Program Outcomes: A Practical Approach*	Provides accountability measurement, helps identify programs needing impact evaluation, and produces tools that allow policymakers, program managers and administrators to collaborate effectively.	Requires a combination of the identification of a business process as well as the linking of relationships using a logic model and sequential outcome indicators.

KEY ANALYTICAL ISSUE # 4: There are Three Types of Inferences that Need to be Drawn for Performance Evaluation			
PERFORMANCE EVALUATION CHARACTERISTIC:	**Tools Used to Identify Indicators**	**Analytical and Methodological Issue Raised**	**Type of Inference**
1. Performance Evaluation assumes that events, activities and outcomes are causally linked	Theory of Change Matrix of Events and Measures	How are activities and events linked to outcomes?	**Relationships**
2. Performance Evaluation Targets End Outcomes 3. Performance Evaluation Indicators use Population-Based Data	Logic Model Identification of Short-term, Intermediate and Long-term Outcomes	What is an outcome? How do indicators measure outcomes?	**Outcomes and Indicators**
4. Performance Evaluation is based on business process 5. Performance Evaluation Tracks Accountability for Managing as Well as Planning and Evaluation	The Results Chain Identification of the business process Policy, program and operational goals	How are processes linked to outcomes?	**Business Process**

KEY ANALYTICAL ISSUE # 5

What You Need to Demonstrate Varies According to the Type of Inference You Wish to Draw

Type	Questions Answered	VALIDITY	RELIABILITY
Relationships	**Is it real?** **How strong is it?** **What direction is it?**	**Demonstration of Counterfactual**	**Demonstration of Universal Relationship**
Outcomes and Indicators	**Are you measuring what you intend to measure?**	**Demonstration of Generalizability**	**Demonstration of Consistency**
Business Process	**Does the process add value in an efficient, well-managed way?**	**Demonstration of Core/Strategic Process**	**Demonstration of Defined Process**

KEY ANALYTICAL ISSUE # 6

Methods and Criteria Used to Demonstrate Validity and Reliability Are Well-Defined According to the Type of Data Inference Being Justified

Type of Data Inference	VALIDITY		RELIABILITY	
	Methods	**Criteria**	**Methods**	**Criteria**
Relationships	**Two approaches:** 1. Research Design 2. Analytical Techniques	**Counterfactual** ▶ Hypothesis Testing ▶ In design choice, *Internal Validity* is preferred over *External Validity* ▶ Desirable to balance Type 1 and Type 2 errors	**Three approaches** 1. Multiple Studies 2. Methodological Triangulation 3. Meta- Analysis Techniques	**Universal Relationship** • Across Time • Different Studies • Different Researchers • Different Methods • Different Settings • Different Populations
Outcomes and Indicators	**Two Requirements:** 1. Population Data 2. Consensus	**Generalizability** ▶ To populations of interest (via census or sampling) **5 Levels** • Face • Content • Predictive • Concurrent • Construct	**Three Aspects** 1. Equivalence 2. Stability 3. Homogeneity **4 Tests** • Parallel/Alternate Forms Test • Inter-Rater Test • Test-ReTest • Split Half Test *(Cronbach's Alpha)*	**Consistency** ▶ Multiple measures of same concept give same result. ▶ Desirable to have a scale with known characteristics *(e.g., Likert; Guttman; Thurstone)* or known statistical properties *(e.g., multidimensional scaling).*
Business Process	**Three approaches:** 1. Strategic Management 2. Process Management 3. Process Improvement	**Core/Strategic Process** ▶ Link to Strategic Plan ▶ Link to Management Expectations ▶ Ability to Benchmark	**Statistical Process Control** Run - ReRun	**Defined Process = (Level 3)** **5 Levels of a Process** 1 = ad hoc, chaotic 2 = repeatable 3 = defined 4 = managed 5 = optimized