The Status of Evaluation in the Federal Government: The Shape of Things to Come?

2nd Annual Environmental Evaluators' Networking Forum

Kathryn E. Newcomer, Ph.D.
The George Washington University

June 14, 2007

Session Objectives

- Discuss the current environment for program evaluation and performance measurement in government and in the nonprofit sector
- Identify some unintended consequences of programmatic evaluation and measurement

Program Evaluation is defined as:

The application of systematic analytical (social science research) methods to address questions about program operations and results.

or

Measurement plus Judgment!!

Performance Measurement is defined as:

- The routine measurement of program inputs, outputs, intermediate outcomes or longer-term outcomes attributed to a program.
- or
- Measurement plus Judgment!

Why evaluate programs???

- For program improvement/development
- For accountability to funders, sponsors
- For knowledge (theory) creation

Theory Underlying Program Evaluation Practice

- Evaluation and/or programmatic performance measurement of programs should be undertaken in order to improve the programs and their outcomes -- through providing useful and timely information about programs.
- So what about the use of the information for the exercise of accountability?

Current Drivers of Evaluation Practice in the U.S.

- Government
  - The Government Performance and Results Act and OMB's PART process at the federal level
  - "Managing for Results" initiatives in states and cities
- United Way
- Foundations
- Boards of directors of nonprofits
- Professionalized staff
- Evidence-Based Policy Movement
- Other Donors

PART Focus on Program Results

- PART stands for Program Assessment Rating Tool
- A set of about 30 questions addressing program design, management and results is to be answered with “Yes, “Large Extent”, “Small Extent,” or “No.”
- The questions include three on achievement of performance goals, one comparing program to other programs with similar purpose and goals, and one on effectiveness.

PART: Old wine in new bottles?

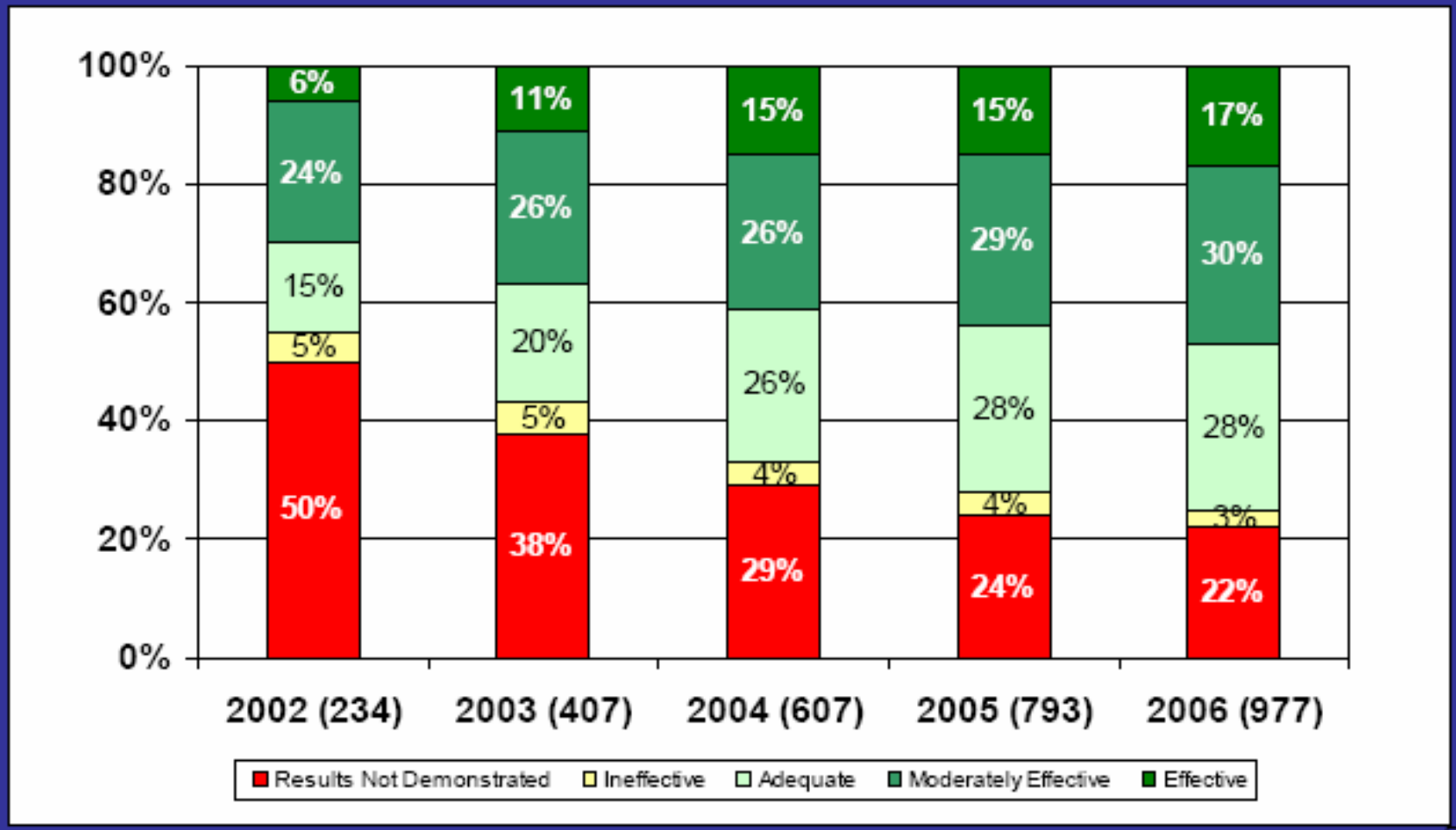
- The focus on program effectiveness by OMB examiners is not new
- What is new?
  - Transparency
  - Explicit quantitative assessments
  - Raising hard questions about the need for managerial and even legislative changes in program design
  - Explicit public attention to the need for rigorous methods to ascertain impact (The RCT push!), and more “hard evidence”

Key PART Question on Program Effectiveness

- Do **independent** evaluations of sufficient **scope** and **quality** indicate that the program is **effective** and achieving **results**? (question 4.5)

Where We Are Today

Distribution of Cumulative Ratings 2002 - 2006



Discussion of PART

- The PART process operates in a political environment--
  - Why is it not surprising that 28% (50% first year to 22% in 2006) of the programs parted thus far were deemed " results not demonstrated?" i.e., failed to reject the null.
  - Why might it be difficult to conduct a mega-PART on a number of programs with seemingly similar goals that are offered in different federal agencies?
  - Why has the PART process not gained widespread Congressional buy-in to use in their budgetary deliberations?

What do the Agency Managers think about the PART Process?

- Initial skepticism
- Concern about inter-rater reliability (in giving PART scores)
- Concern about what constitutes rigorous methods and “hard evidence”
- Confusion on what exactly constitute “**independent** evaluations of **sufficient scope** and **quality**”

Pressures on Public and Nonprofit Managers to Measure Program Performance

Facilitating:

Internal Factors:

- Executive Branch Initiatives
  - Budget calls for non-financial performance measures
  - President's Management Agenda 2001-present
- Legislation
  - Laws affecting all programs, e.g. Government Performance and Results Act
  - Laws requiring performance measures for specific programs

Pressures from Environment:

- Citizens Demands for Evidence of Program Results
- Evidence-based Policy Movement
- Success stories from other Jurisdictions and other Countries
- Accounting Profession Use of Performance Auditing

Inhibiting:

Internal Factors:

- Insufficient Authority and/or flexibility to Execute Needed Change
- Mixed Signals from Legislative Committees of Use of Measures in budget Process
- Multiple Calls for Measurement in Different Laws and Executive Directives
- Complex Relationships among service Delivery/Regulatory Partners
- Unclear Expectations about Use Performance Data
- Unclear Expectations about Incentives/Punishment for Performance

Pressures from Environment:

- Citizen Expectations of Clear Evidence of Program Results
- Anxiety about Comparing Performance across Jurisdictions
- Lack of Comparable, Reliable Data Collection across Jurisdictions

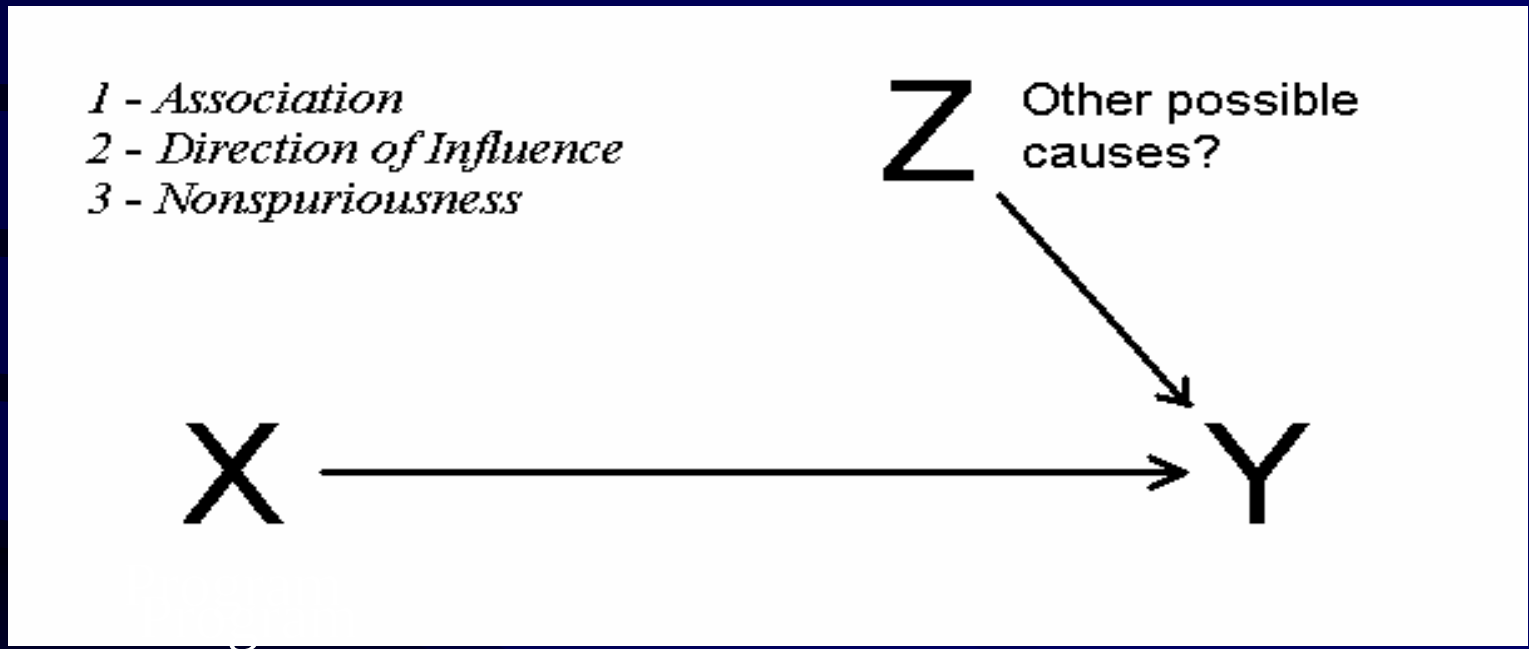
Consequences of Measurement?

- “Do you count what can be counted rather than what counts?” (Einstein)
- Is the rush to measure expanding our capacity or is capacity shaping measurement?
  - Are we adequately auditing validity and reliability of data?
- Are we interpreting the numbers out of context? (any systems thinking?)
- Are calls for “hard evidence of effectiveness” in some areas even within reason?
- What is the impact of setting targets?
  - Threshold effects?
  - Outputs distortion?
- What about rankings?
  - Validity of criteria?
  - Reliability of data used?

And What about the Measurement of Program Results?

- How might we set up adequate comparisons to rule out rival explanations for the results, or impacts, of programs?
- Is construction of counterfactuals even possible for some environmental programs?
- How do we make the case for plausible attribution, or even contribution?

Causal Inference or Plausible Attribution or Contribution?



3 Elements of Causal Inference

1. Temporal order 2. Co-Variation 3. Nonspuriousness

Consequences of Judgment?

- Are GPRA and PART requirements treated as “paperwork exercises”?
- What will OMB do post-Bush’s PMA (PART)?
- Is path dependency (in performance measurement) limiting incremental learning?
- What are the results of shame games?
- Is defensive gaming undercutting risk-taking?
- Is impression management increasing?

Lingering Issues

- The time and resources requirements of completing PART assessments!
- High expectations of “experimental” research (RCTs) to meet PART requirements!
- The resources requirements for completing outcome or impact evaluations!
- The need to meet accountability demands trumping real programmatic learning!