Evaluation Dialogue
Between OMB Staff and Federal Evaluators

Digging a
Bit Deeper into
Evaluation
Science

July 2006

Why are we here today – How can we benefit from this dialogue?

- Obtain clarity for evaluation community on what approaches are appropriate for PART/BPI

- Encourage understanding of evaluation approaches & products generally accepted by evaluators

- Ultimately, we all aim to improve Federal programs, solving problems to increase effectiveness

Format of Dialogue

- Session I: Brief overview of evaluation approaches
- Session II: Examples of evaluation approaches and discussion

What is program evaluation?

- A systematic assessment of how well a program is working
- Consists of various activities:
  - Needs assessment
  - Design assessment
  - Process/Implementation evaluation
  - Evaluability assessment
  - Outcome and Impact evaluations
  - (Formative vs Summative)

How are evaluation questions and types relevant to the PART?

- Needs assessments and process evaluations
  - Primarily relevant to PART Sections 1, 2, & 3
- Outcome and impact evaluations
  - Primarily relevant to PART Section 4 and PART questions 2.6 & 4.5

Why should we conduct evaluations?

Provide feedback for program improvement and external accountability

- Answer evaluation questions about results and the processes that managers directly control to achieve results

- Document effectiveness and value added to society

Evaluation / Management Cycle

Evaluation Feedback

- Feedback Evaluation Findings to Managers

- Refine Program
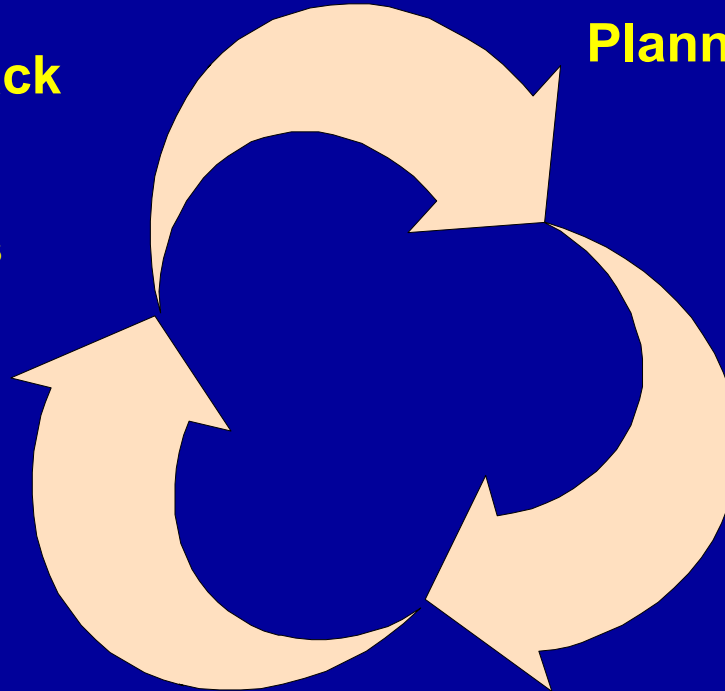
Planning/Decision Making

- Identify Needs, Problems, Solutions, Refinements

- Conceptualize Program

- Formulate Evaluation Questions and Design

Implementation

- Actualize the Program Plan

- Collect Evaluation Data

- Analyze Data

Who conducts evaluations?

- Professionals blend a wealth of scientific approaches and perspectives

- Within federal agencies, evaluators are found in a variety of offices

- Field is supported by professional organizations and degree programs

(See evaluation information resources handout)
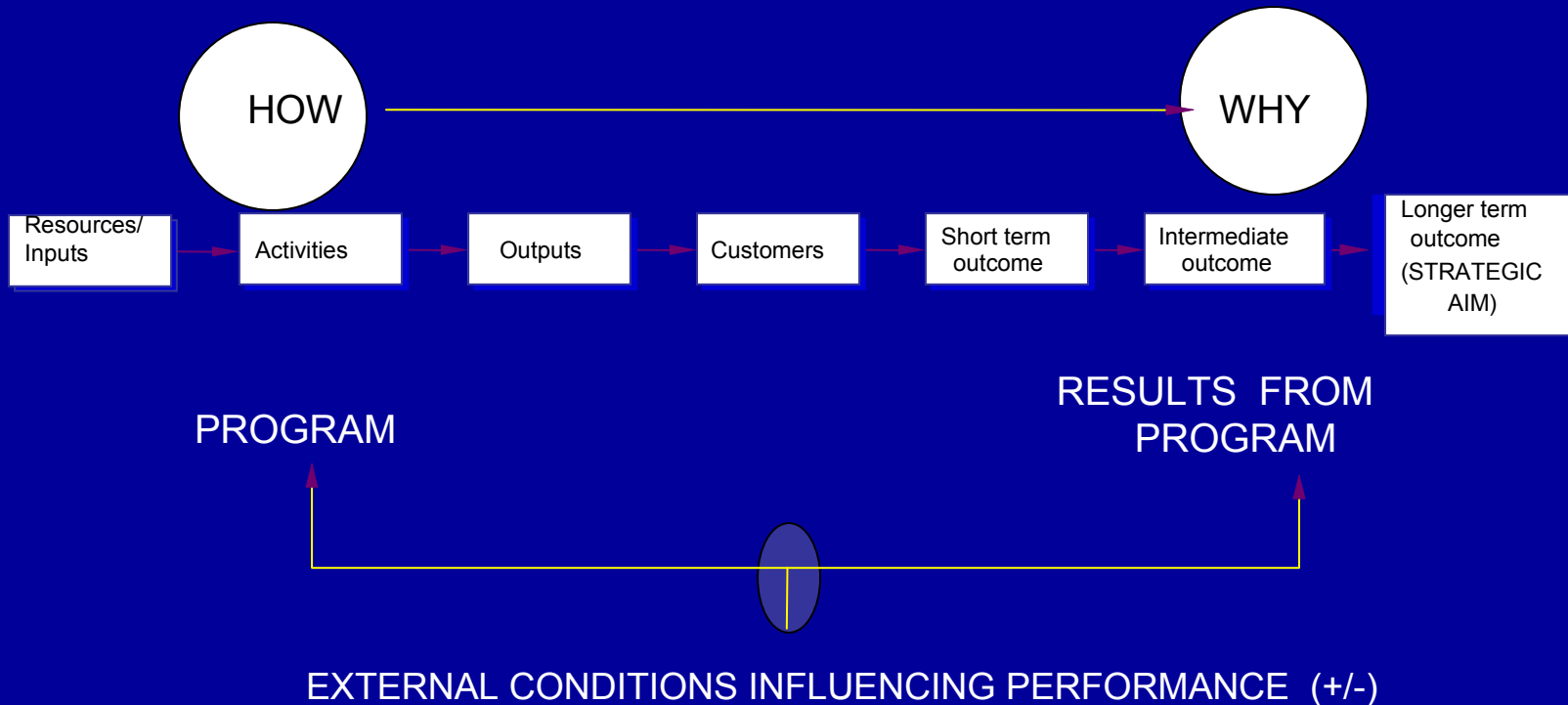
What steps do evaluators use?

1. **Conceptualize the program**
2. **Develop relevant and useful evaluation questions**
3. **Select appropriate evaluation approaches for each evaluation question**

4. **Collect data to answer evaluation questions**
5. **Analyze the data and draw conclusions**
6. **Communicate results and recommendations**

Step 1. Conceptualize the Program
by showing simple flow of logic

The Logic Model

"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."
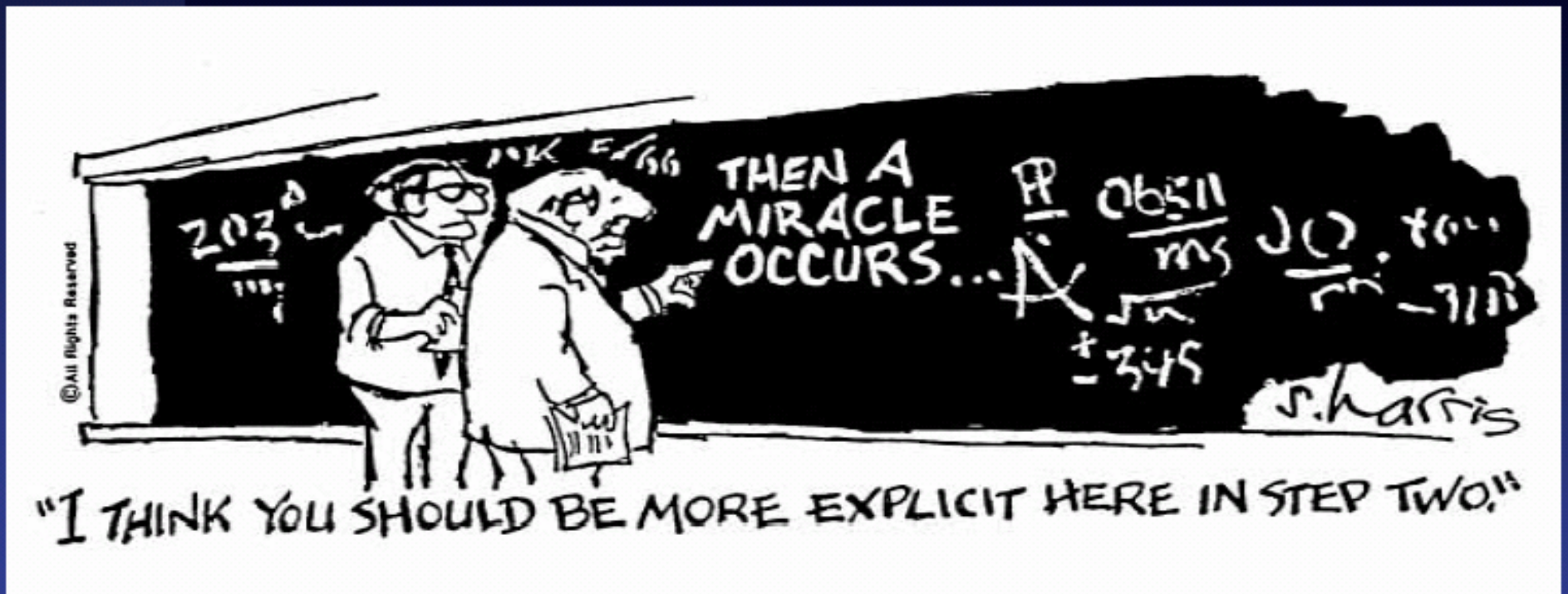
Logic models illustrate the causal relationships among program elements and define program success

HOW

WHY

| Resources/ Inputs | | Activities | | Outputs | | Customers | | Short term outcome | | Intermediate outcome | | Longer term outcome (STRATEGIC AIM) |

PROGRAM

RESULTS FROM PROGRAM

EXTERNAL CONDITIONS INFLUENCING PERFORMANCE (+/-)

Generic Logic Model Worksheet

Inputs	Outputs		Outcomes-Impact		
	Activities	Participation	Short term	Medium Term	Long Term

Situation
Needs and Assets
Symptoms versus problems
Stakeholder engagement

Priorities:
Consider:
Mission
Vision
Values
Mandates
Resources
Local Dynamics
Collaborators
Competitors
Intended Outcomes

What we invest
Staff
Volunteers
Time
Money
Research base
Materials
Equipment
Technology
Partners

What we do

Conduct workshops, meetings
Deliver services
Develop products, curriculum, resources
Train
Provide counseling
Assess
Facilitate
Partner
Work with Media

Who we reach
Participants
Clients
Agencies
Decision-makers
Customers

Satisfaction

What the short term results are
Learning
Awareness
Knowledge
Attitudes
Skills
Opinions
Aspirations
Motivations

What the medium term results are
Action
Behavior
Practice
Decision-making
Policies
Social Action

What the ultimate impact(s) is
Conditions
Social
Economic
Civic
Environmental

Assumptions

External Factors

Evaluation
Focus - Collect Data – Analyze and Interpret - Report

Step 2. Develop relevant and useful evaluation questions

Why are good questions important?

- Articulate the issues and concerns of stakeholders
- Posit how the program is expected to work and its intended achievements
- Frame the scope of the assessment
- Drive the evaluation design

Table 1: Common Evaluation Questions Asked at Different Stages of Program Development

Program Stage	Type of Activity	Common Evaluation Questions
Program design	Needs assessment	•What are the dimensions of the problem and the resources available to address it?
	Design assessment	•Is the design of the program well formulated, feasible, and likely to achieve the intended goals?
Early stage of program or new initiative within a program	Process evaluation or implementation assessment	•Is the program being delivered as intended to the targeted recipients? •Is the program well managed? •What progress has been made in implementing new provisions?
Mature, stable program with well-defined program model	Evaluability assessment	•Is the program ready for an outcome or impact evaluation?
	Outcome monitoring or evaluation	•Are desired program outcomes obtained? •Did the program produce unintended side-effects?
	Process evaluation	•Why is a program no longer obtaining desired outcomes?
	Net impact evaluation	•Did the program cause the desired impact? •Is one approach more effective than another in obtaining the desired outcomes?

Step 3. Select appropriate evaluation approaches to answer evaluation questions

How do we control for alternative explanations of effects?

- Ensure conditions necessary for establishing causality

- Use design elements that control for alternative explanations

- Use multiple indicators

- Build strong argument

What are criteria for selecting an evaluation design?

- Matches evaluation question
- Fits available resources
  - Time and Funds
- Data are available/ Can be acquired
- Appropriate to the program type
  - Regulatory, Research, Service Delivery

Process and Outcome Monitoring or Evaluation

Compares program performance to a pre-existing goal or standard, for example:

- OMB R&D criteria of relevance, quality, and performance
- productivity, cost effectiveness, and efficiency standards
- customer expectations or industry benchmarks

Typically used with research, enforcement, information and statistical programs, business-like enterprises, and mature, ongoing programs with:

- complete national coverage
- few, if any, alternative explanations for observed outcomes

Example of Outcome Monitoring: Mediterranean Fruit Fly Program

- Question: Is the program controlling the "Medfly" population at the desired target level?

- Outcome data: Weekly monitoring of the "Medfly" population level and dispersion, to detect outbreaks

- Evaluation Design: Review program policies, practices, and resources to identify causes of outbreaks

Quasi-Experimental Single-Group Design

Compares outcomes for program participants before and after the intervention:

- Multiple data points are collected over time
- Statistical adjustments or modeling control for alternative causal explanations

Typically used with regulatory and other programs where:

- clearly defined interventions have distinct starting times
- coverage is national, complete
- random assignment of program participation is NOT feasible, practical, or ethical

Example of Quasi-Experimental Single-Group Design: Baby Walker

- Question: Has the safety standard been effective in reducing injuries?

- Evaluation Design: Interrupted time-series compared injury rates before and after introduction of regulatory standard

- Controlled for alternative explanations through measurement and logical elimination of possible alternatives identified

Baby Walker-Related Injury Rate: 1981 to 2001
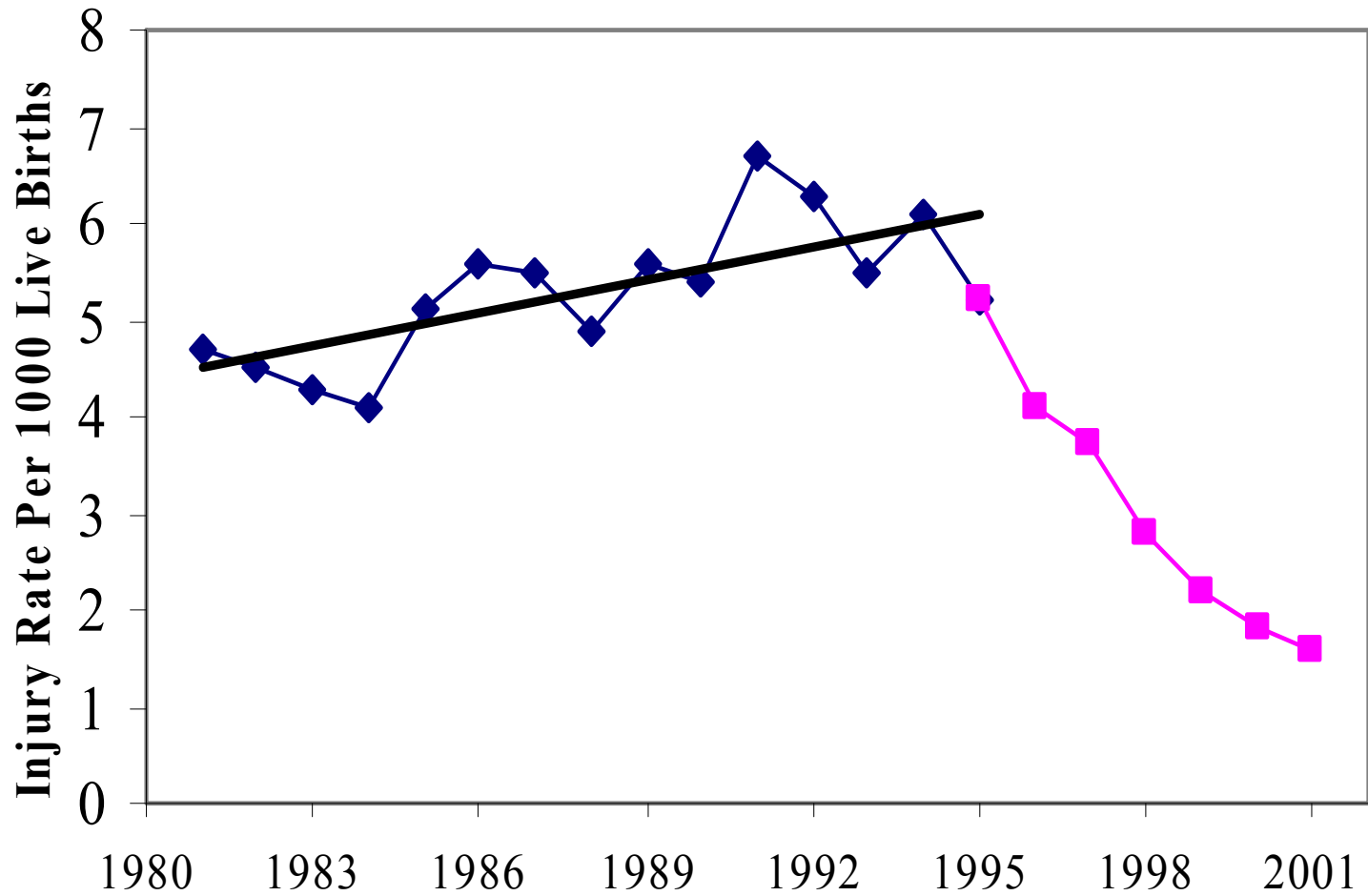
Quasi-Experimental Comparison-Group Design

Compares outcomes for program participants with outcomes for a comparison group selected to closely match participants on key characteristics:

- Key characteristics are plausible alternative explanations for a difference in outcomes
- Outcomes are measured before and after the intervention

Typically used for service and other programs where:

- clearly defined interventions can be standardized and controlled
- coverage is limited
- random assignment of participants is NOT feasible, practical, or ethical

Example of Quasi-Experimental Comparison-Group Design: GI Bill

- Question: Did educational assistance meet needs of beneficiaries (veterans)?
- Evaluation Design: Compared program users with non-users on education achievement, income attainment, and career goals
- Statistically controlled for differences in demographic characteristics, educational level, and military rank

Randomized Experiment Control-Group Design

Compares outcomes for those randomly assigned to participate ("treatment" group) with outcomes for those who did not participate ("control" group):

- Outcomes are measured before and after the intervention

Typically used for service and other programs where:

- clearly defined interventions can be standardized and controlled
- coverage is limited
- random assignment of participants is feasible and ethical

Example of Randomized Design: Upward Bound

- Question: Does the program help low income, academically high-risk students complete high school and attend college?
- Evaluation Design: Applicants were randomly selected to the program and compared to non-selected applicants
- Random assignment controlled for many alternative explanations, such as demographics and motivation level

Table 2: Common Evaluation Approaches For Assessing Program Effectiveness

Typical designs used to assess program effectiveness	Design features that help control for alternative explanations	Best suited for (typical examples)
Process and outcome monitoring or evaluation	Compares performance to a pre-existing goal or standard. For example: • OMB R&D criteria of relevance, quality, and performance • Productivity, cost effectiveness, and efficiency standards • Customer expectations or industry benchmarks	Research, enforcement, information and statistical programs, business-like enterprises, and mature, ongoing programs where: • Coverage is national, complete • There are few, if any, alternative explanations for observed outcomes
Quasi-experiments – Single Group	Compares outcomes for program participants before and after the intervention. • Outcome data are collected over multiple points in time • Statistical adjustments or modeling control for alternative causal explanations	Regulatory and other programs where: • Clearly defined interventions have distinct starting times • Coverage is national, complete • Random assignment of participants is NOT feasible, practical, or ethical
Quasi-experiments – Comparison Groups	Compares outcomes for program participants with outcomes for a comparison group selected to closely match participants on key characteristics. • Key characteristics are plausible alternative explanations for a difference in outcomes • Outcomes are measured before and after the intervention (pretest, posttest)	Service and other programs where: • Clearly defined interventions can be standardized and controlled • Coverage is limited • Random assignment of participants is NOT feasible, practical, or ethical
Randomized experiments – Control Groups	Compares outcomes for those randomly assigned to participate ("treatment group") with outcomes for those assigned not to participate ("control" group) • Outcomes are measured before and after the intervention (pretest, posttest)	Service and other programs where: • Clearly defined interventions can be standardized and controlled • Coverage is limited • Random assignment of participants is feasible and ethical

How do we determine the quality of an evaluation?

- Evaluation questions have been answered fully
- Findings support conclusions
- Conclusions portray strong causal arguments
- Study meets professional evaluation standards
  - Utility, Feasibility, Propriety, and Accuracy

Checklist of Questions for Assessing the Quality and Usefulness of a Program Evaluation

Are the study's objectives stated?

Were the objectives appropriate with respect to the developmental stage of the program?

Is the study design clear?

Was the design appropriate given the study objectives?

Was the indicated design in fact executed?

Did the variables measured relate to and adequately translate to the study objectives and are they appropriate to the study objectives and are they appropriate for answering the client's questions?

Are sampling procedures and the study sample sufficiently described? Were they adequate?

Are sampling procedures such that policymakers can generalize to other persons, settings, and times of interest to them?

Is an analysis plan presented and is it appropriate?

Were data-collector selection and training adequate?

Were there procedures to ensure reliability across data collectors?

Were there any inadequacies in data collection procedures?

Were problems encountered during data collection that affect data quality?

Are the statistical procedures well specified and appropriate to the task?

Are the conclusions supported by the data and the analysis?

Are study limitations identified?

What possibly confounds the interpretation of the study findings?

How can we work together to ensure the best evaluations?

- Develop a common understanding of the program via logic model and/or strategic plan
- Develop good evaluation questions
- Select appropriate evaluation study designs to answer questions
- Draw on program conceptualization to identify needed performance measures
- Develop multi-year plan to meet evaluation information needs

Federal Evaluation Leaders

Working with OMB to dig up the best evaluation information possible!

Contributor Acknowledgements

- LCdr Eric Bernholz, USCG
- Joseph Carra, DOT
- Patrick Clark, DOJ
- Alan Ginsburg, ED
- Marcelle Habibion, VA
- John Heffelfinger, EPA

- David Introcaso, HHS
- Cheryl Oros, HHS
- N.J. Scheers, CPSC
- Stephanie Shipman, GAO
- Linda Stinson, DOL
- Bill Valdez, DOE