



Measuring Science



Presentation to the Smithsonian
Institution Science Division

July 12, 2007

Office of Policy and Analysis



Two questions

- How do science research organizations in various sectors measure quality, relevance, and performance of research programs?
- What criteria or framework do they use to inform decision-making with respect to research direction and support of science programs?
- Methods
 - Review of the literature on measurement of science R&D
 - Interviews with representatives of science research organizations (universities, museum-based and other research institutes, federal science agencies)

Interviewees

- **Ann Arvin**, Dean of Research, Stanford University
- **Carl O. Bostrom**, retired Director of Applied Physics Lab, John Hopkins University
- **Kathleen Buckley**, Associate Provost for Science, Harvard University
- **Beth Burnside**, Vice Chancellor for Research and Professor of Molecular and Cell Biology, University of California, Berkeley
- **Marie Colton**, Director of National Ocean Service, NOAA
- **Jim Colvard**, former Associate Director of Applied Physics Lab, former Deputy Director, Office of Personnel Management, former Deputy Chief, Naval Material Command
- **Susan Cozzens**, Professor of Public Policy and Director of Technology Policy and Assessment Center, Georgia Institute of Technology; member National Academies Committee on Science, Engineering and Public Policy (COSEPUP)
- **Sharon D. Drumm**, Staff Officer, National Program Staff, USDA - Agricultural Research Service
- **Steven J. Fluharty**, Associate Vice Provost for Research, University of Pennsylvania
- **Darrel Frost**, Associate Dean of Science for Collections and Curator, Division of Vertebrate Zoology (Herpetology), American Museum of Natural History
- **Lance Grande**, Senior Vice President and Head of Collections and Research, Field Museum
- **Elliot Hirshman**, Chief Research Officer and Professor of Psychology, George Washington University
- **Jim Keeley**, Associate Director of Communications, Howard Hughes Medical Institute
- **Ronald N. Kostoff**, Director of Technical Assessment, Office of Naval Research
- **Theodore (Ted) Poehler**, Vice Provost for Research, Johns Hopkins University
- **Eva J. Pell**, John and Nancy Steimer Professor of Agricultural Sciences, Vice President for Research and Dean of the Graduate School, Penn State University
- **Gregory Tassej**, Senior Economist, NIST
- **William J. (Bill) Valdez**, Director of Office of Workforce Development for Teachers and Scientists, Department of Energy, Office of Science

Outline

- Context for support of basic research
- Approaches to science measurement
- Current Practices: Universities and Research Institutes
- Current Practices: Federal Science Programs
- Key Points with Implication for SI Science

Context for Support of Basic Research



The Post-War “Social Contract”

- 1946: Vannevar Bush’s report to President Truman (*Science: The Endless Frontier*) lays down terms for a new “social contract” between federal government and national scientific community
- Active, large-scale government support for science is seen as crucial—provides foundation for new technologies that enhance national strength (ie, defense) and welfare (ie, health care)

The “Golden Age”

Next 3-4 decades are “Golden Age” of U.S. science

- Federal government provides big pot of \$\$\$ in areas of specific interest (defense, energy, health, space, environment, etc.) as well as “blue-sky” science (at universities), with few strings attached
- Scientific community itself determines how \$\$\$ are allocated among specific research programs, with little input or oversight from other stakeholders
- Steady stream of technological wonders—with benefits accruing first and foremost to U.S. firms and citizens—suggests the model works.

The End of the “Golden Age”

Starting in 1970s, “social contract” comes under strain:

- Traditional models of commercial diffusion of scientific knowledge (“pipeline,” “spin-off”) are called into question.
- Increased global competition raises question of whether economic benefits of government-funded scientific research necessarily go to U.S. firms
- End of Cold War reduces concerns about “national security” dimension of scientific leadership
- General concerns about accountability and efficiency in the federal government

The Accountability Movement

- Gaining momentum in early 1990s, the performance movement calls for greater efficiency and accountability from not-for-profit organizations and taxpayer-funded government agencies
 - Government Performance and Results Act (1993)
 - President's Management Agenda (2001)
 - Budget & Performance Integration
 - Program Assessment Rating Tool (FY 2003 Budget)

GPRA (1993)

- Aims to “...improve Federal program effectiveness and public accountability by promoting a new focus on **results...**”
- Requires federal agencies to
 - Develop strategic plans and annual performance plans and reports
 - Identify output and outcome measures for all programs

Program Assessment Rating Tool (PART)

- Examines various factors that contribute to program **effectiveness**
- Assesses if and how **evaluation** is used to inform program planning and to corroborate program results
- Leads to development and institutionalization of validated, systematic approaches to program management and performance evaluation

Four Parts of PART

- Program Purpose and Design (20% of score)
 - *To assess whether program design and purpose are clear and defensible*
- Strategic Planning (10% of score)
 - *To assess whether the agency sets valid annual and long-term goals for the program*
- Program Management (20% of score)
 - *To rate agency management of the program, including financial oversight and program improvement efforts*
- Program Results (50% of score)
 - *To rate program performance on goals reviewed in the Strategic Planning Section and through other evaluations*

President's Management Agenda: "The Scorecard" (2001)

- Budget & Performance Integration Initiative builds on GPRA
 - Strategic plans contain limited number of outcome-oriented goals and objectives
 - Individual performance appraisal plans link to mission, goals, and outcomes
 - Full cost of achieving performance goals to be reported; marginal cost analysis
 - Efficiency measure for all PARTed programs
 - Use of PART evaluations to direct program improvements; justify funding requests

Science Community Concerns

- Basic research is difficult to measure
 - Likely outcomes not calculable in advance – results often serendipitous
 - Knowledge gained not always of immediate value or application
 - High percentage of negative determinations or findings
- Requiring identification of outcomes will
 - Lead to “short-termism”, i.e., encourage agencies to measure what is easy and neglect what is important
 - Discourage high-risk performance and stifle creativity
- How to define “program” (by budget, organizationally, by field, by initiative...)

COSEPUP on GPRA (1999)

- National Academies Committee on Science, Engineering, and Public Policy
 - Useful outcomes of basic research cannot be measured directly on an annual basis—usefulness of new basic knowledge is inherently unpredictable and must be measured by historical reviews based on long timeframe
 - Both applied and basic research programs can be evaluated meaningfully on a regular basis
 - One size does not fit all. Measurement needs to match the character of the research—different timescales; what is measurable and what is not
 - The most effective means of evaluating federally funded research is expert review including quality review, relevance review, and leadership benchmarking

OMB–R&D Investment Criteria Now Incorporated in the PART

- Program **relevance, quality, and performance**
- Intended to address full cycle of planning, management, prospective assessment, and retrospective review of whether investments were well-directed, efficient, and productive
- Not intended to “predict the unpredictable” but to improve management of research programs – “Vague goals lead to perpetual programs achieving poor results”

Program Relevance

- R&D investments must have clear plans, must be relevant to national priorities, agency missions, relevant fields, and “customer” needs, and must justify their claim on taxpayer resources
 - Theoretical significance (Enriched the field through insights? Developed concepts methods or models that apply widely?)
 - Mission relevance (relevance of knowledge produced to practical goal of the program)

Program Quality

- Programs should maximize the quality of the R&D they fund through use of a clearly stated, defensible method for awarding a significant majority of their funding, i.e., a competitive, merit-based process
- Programs must assess and report on the quality of current and past R&D, i.e., benchmarking internationally or across agencies

Program Performance

- R&D programs should maintain a set of high priority, multi-year objectives with annual performance outputs and milestones that show how one or more outcomes will be achieved
- Metrics should be defined to encourage individual performance, but also to promote broader goals such as innovation, cooperation, education, and dissemination of knowledge, applications, or tools

The Upshot

- “Then”: Working assumption was that government support for science automatically advances U.S. national security and welfare
 - “What’s good for science is good for the nation”
- “Now”: Increasing pressure to *demonstrate* results that benefit taxpayers
 - Managing science vis-a-vis GPRA is seen by many as prerequisite to federal funding

Approaches to Science Measurement



Different “Cuts” for Evaluation

- By stage of research, i.e., distance of research activities from practical application: basic/ fundamental, applied, development
- By unit of measure: individual, project, program, portfolio, organization, system

Evaluation Method by Stage of Research



Evaluation Methods:

- \$ grants
- Scientists trained
- Publication counts
- Citation analysis
- Expert judgement



Evaluation Methods :

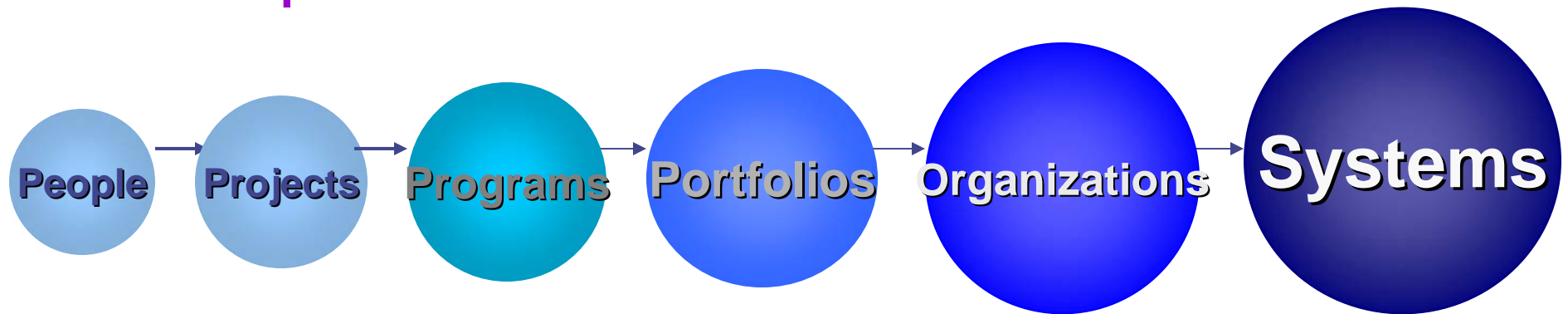
- \$ grants
- Publication counts
- Patents
- Expert judgement



Evaluation Methods:

- ROI
- Economic growth
- Global competitiveness

Evaluation Method by Size of Science Enterprise



Evaluation Methods:

- Peer Review
- Output Metrics
- Input metrics

Evaluation Methods:

- Input metrics
- Peer Review
- Output Metrics (incl. cost, schedule, technical milestones)

Evaluation Methods:

- Committees of Visitors
- Output & Outcome Metrics
- Case Studies
- Randomized Trials

Evaluation Methods:

- Advisory Committee reviews
- Econometric Modeling
- Risk/Options Modeling
- Case Studies

Evaluation Methods:

- National Academy reviews
- Econometric Modeling
- Committee Reviews
- Case Studies

Evaluation Methods:

- NAS/COSEPUP International Benchmarking
- Longitudinal Studies
- Innovation Indexes
- Case Studies
- Network Analysis
- Modeling

Adapted from DOE Office of Science presentation: Valdez, 2005

Research Performance – Logic Model

Golden Lion Tamarin project



Evaluation Results:

- Reintroduction of captive born GLTs
- Increase of GLTs living in wild
- Training of Brazilian scientists
- Education of locals
- Understanding of forest ecology

Center for Conservation & Evolutionary Genetics

Evaluation Results:

- New knowledge of genetic management of wild and captive populations
- Development of software and database tools
- Trained conservation biologists

NZP



Evaluation Results:

- New knowledge that assists survival and recovery of species and habitats
- Trained scientists

Smithsonian Science



Evaluation Results:

- Resources focused in four areas of scientific inquiry
- Documented advances in understanding the origin and nature of the universe; formation and evolution of earth and similar planets; biological diversity; and human diversity and cultural change

U.S. R&D



Evaluation Results:

- Increased Quality of Life
- New Knowledge
- Healthy Environment
- New Technologies
- Understanding of Complex Biological Systems
- Increased National Security

Types of Measures

- Inputs
- Outputs
- Efficiency
- Outcomes

- **BUT numbers will be misused – methods not useful without a framework within which to place them**
 - What is to be measured?
 - Against which yardstick?
 - For whose benefit is the measurement being made?

Inputs

- Funding (government, industry, academic)
 - Amount of grant money brought in
- Human resources (scientists, engineers, other staff; post-graduate students)
- Infrastructure (advanced instrumentation, equipment, labs, etc.)

Outputs – Publication Counts

- Indicator of *quantity* of knowledge produced (not quality)
- No standard value across fields, i.e., chemists publish numerous short papers, mathematicians publish infrequently but in depth; geologists publish massive accounts of field work
- Different weights of books, articles, co-authored articles, etc.
- May skew numbers upward as researchers respond to “reward system”
- Boundaries are set to get at quality, e.g., peer-reviewed journals or high impact journals in the field

Outputs: Citation Counts

- Indicator of *quality* of research, i.e., influence and transfer of knowledge – “an unobtrusive form of wide-scale peer review”
- Citation patterns differ by field of research – biochemists average 30 references per article compared to 10 for mathematicians
- Not all types of papers cited at equal rate, i.e., evidence that methodological papers cited more often than others; experimental work more often than theoretical
- Peculiar effect of negative citations

Outputs: Advanced Bibliometrics

- Co-citation analysis: pairs or groups of articles cited together in other articles
- Co-word analysis: linking papers by keywords or sets of words
- Scientific mapping: “map” or model of literature output of scientific fields
 - Indicators of centers of scientific excellence; intellectual connections between organizations; linkages between subject areas

Outputs: Awards and Honorific Positions

- Indicator of *quality of researchers* supported by a program, and thus indirectly of the quality of knowledge produced
- Similar kinds of quality measures
 - Papers accepted for presentation at national and international conferences
 - Roles of investigators in field of research, e.g., journal editors, conference organizers, invited speakers
 - Collaborative programs with outside scholars

Outputs: Patents, Licenses, New Technologies

- Indicator of connection between research and the marketplace – more useful in applied areas where technology is primary output
- Less relevant to basic science

Outcomes: Expert Judgment

- Most widely used approach in research evaluation
- ***Prospective*** peer review panels that judge proposals
- ***Retrospective*** external panels that judge quality of projects supported
- Essentially a subjective process – results are highly dependent on choice of reviewers

Outcomes: User/Stakeholder Evaluations

- Retrospective evaluation by “customers” or next-stage users in areas of practice that will benefit from knowledge
 - Surveys
- Prospective involvement in shaping of programs

Economic Approaches



Historically, most efforts to rigorously measure outcomes of scientific research have come from the field of economics.

- Some controversy about appropriateness of economic tools
- (But other fields lack standardized, quantitative welfare outcome measures.)

The Big Picture: Public Goods

- Economics provides compelling argument for government support of basic science: It is a *public good*.
 - By definition: A good (service, activity, etc.) that will not be supplied in adequate quantities by the private sector, because benefits cannot be captured by private-sector suppliers.

Basic Science as a Public Good

- Long-term benefits accrue to *third parties* in *unpredictable ways* over *long periods* of time.
 - Classic example: Boolean algebra, the obscure 19th century intellectual curiosity that became the formal basis of modern computing.
 - In a sense, long-run economic return to this highly abstract intellectual work could be reckoned in the trillions of dollars.
 - *How much of it accrued to George Boole and his funders?* Precisely **\$0,000,000,000,000.00**.
 - *Would this work have been undertaken if the motivations behind it were private economic gain...?*

Use of Economic Analyses

- Economic analyses are best for assessing applied science and technology programs that result in products and processes that either
 - Carry market values, or
 - Have close market analogs that can be used to estimate imputed dollar values
- However...
 - Economists like to think *everything* can be valued in terms of \$\$\$, if you are clever enough.
- Example: Field of environmental economics
 - Regularly assigns dollar values to non-market “products and processes” (clean air, etc.) that have no close market analogs.

Who Cares?

Is all this of any interest to SI science?

Possibly...

Example: Valuing Biodiversity

- Preservation of biodiversity is an important part of SI's mission.
 - Imagine a future when SI is asked what this is “worth” to the public, given all the other pressing demands on federal funds
 - How could economic approaches help make the case?

Case Study Approach

- Pick some endangered charismatic megafauna at the Zoo—Say, *pandas*...?
- Collect survey and market data:
 - How much do visitors pay to travel to the Zoo to see Tai Shan? (Market proxy)
 - How much do they pay for panda toys and souvenirs? (Market proxy)
 - How much do they say it is worth to them just to know that these creatures will not die out in the wild? (Contingent valuation)
 - Etc.

Comprehensive Approach

(Requires a substantial budget)

- Create a well-defined index of global (or national or regional) biodiversity
- Investigate the statistical correlation between changes in that index and changes in well-defined indicators of social welfare
- If desired: assign dollar values to indicators of social welfare (economic valuations for human health, life, etc. are available)

Caveat

- Due to unpredictable nature of returns to basic research, economic approach is better suited for *retrospective* evaluation.
 - (For applied research to develop some practical product or process, prospective economic analysis may be feasible. May apply to some SI research.)
- Thus:
 - Economic analysis is better at providing historical evidence that some type of research has “paid off” (in an economic sense) in the past.
 - It is not so good at providing evidence that some type of scientific research will produce economic value in the future.

Beyond Economics: Psychic Value of Public Goods



Photo Credit: By Carol T. Powers -- Bloomberg News

U.S. Declares Bald Eagles No Longer Threatened

By David A. Fahrenthold
Washington Post Staff Writer
Friday, June 29, 2007; A03


The bald eagle was removed from the federal list of threatened and endangered species yesterday during a ceremony at the [Jefferson Memorial](#), as officials and environmentalists celebrated the national symbol's historic recovery.

"Today, I'm proud to announce that the eagle has returned," [Interior Secretary Dirk Kempthorne](#) said in a conference call with reporters yesterday afternoon.


During the ceremony yesterday morning, Kempthorne signed paperwork "delisting" the eagle, which in 40 years has rebounded from 417 breeding pairs in the continental United States to about 10,000.



Current Practices: Universities and Research Institutes



Harvard, UC Berkeley, Stanford
Johns Hopkins, George Washington
Penn State, U-Penn
AMNP, Field, HHMI



Individual Accomplishments Looked at in Tenure, Post-tenure and Annual Reviews

- Dominant method of review for tenure is academic committees of peers
 - Described variously as “harsh set of standards” (JHU), “elaborate” (Berkeley), and “prolonged” (Stanford)
- Peer reviews consider:
 - Scholarship: publications, presentations
 - Ability to bring in extramural funding
 - Ability to build a respectable program
 - Reputation as determined by leading scholars (letters of recommendation)
 - Leadership (organizing symposia; editorial boards)

Variations on Performance Metrics

- Stanford: “Letters of recommendation are basically the only things that really matter when reviewing a candidate’s abilities. Other metrics of performance like publications and citations do not really identify quality”
- Penn State: three criteria – teaching and learning, research and creative accomplishments, and service (active role in the community)
- Berkeley: emphasis on collaborative, multi-disciplinary enterprise
- JHU: emphasis on practical utilization of research for benefit of public
- Harvard and U-Penn look at dollars brought in/square feet of space

External Review

- Some universities have external reviews of all departments; others only certain ones (owing to decentralization). Reviews range from five-year cycle to every decade and for some departments it is tied to accreditation.
- Visiting Committees give frank advice on how well a department is doing and provide justification for internal changes (Berkeley)

Internal Review Groups

- Berkeley – Budget Committee – critically appraises faculty coming up for review; provides more distance than supervisors and has the excellence of Berkeley as its charge
- Harvard – University Planning Committee for Science and Engineering – provides a more holistic view and coordinates research areas across spectrum of science programs

New Research Directions

- Universities tend to be decentralized; schools have great autonomy and new program directions are generally faculty-driven
 - Schools make case for hiring, new facilities/equipment
 - Know relevant areas of study from attendance at Science Academy events (GWU)
 - “General consensus in the world that certain areas of research are vital, e.g., cell engineering” (JHU)
- Outside review committees suggest new directions (Stanford)
- Appointment of regent professors (GWU); visiting scholars (Stanford) who are expert in new research field
- Strategic planning at high level for programs that are interdisciplinary and/or cut across research areas (JHU)

Shutting Down a Line of Research


- In general, decisions made at unit level
 - sense that it is not important
 - can't develop funding or otherwise be sustained
- GWU has shut down programs (e.g., Dental School) due to lack of funds, lack of student demand and/or low level of research productivity
- Penn State Board determines which science projects are strong and which are weak and can be phased out during strategic planning process (every 5 years)

American Museum of Natural History and Field Museum

- AMNP: Two required criteria for grant of tenure to curatorial staff (Research productivity evidenced by peer reviewed scholarly publications and award of competitive extramural research grants) and one optional criteria (professional activities and popularization of science)
- Field: Measures four areas – research, curation, education, and service/administration.
 - Emphasis on collaboration with research universities in area; Field curators mentor 60 resident graduate students and serve as adjunct faculty at surrounding universities

Howard Hughes Medical Institute

- 300 mostly biomedical investigators who are also faculty members at 66 distinguished universities and medical schools
 - Benefits from investigators' collaboration within their university departments
 - HHMI demands creativity and innovation; funds high risk/high reward research that would not be funded by NIH
 - Investigators go through rigorous peer review after 5 years and contract is terminated or renewed for another 5 years
 - **Deletion test**: Would biomedical work in the field be lesser if investigator's work was deleted?
 - New approach this year: open competition for investigators
 - any scientist from list of distinguished institutions can nominate themselves
 - New research direction: Janelia Farm, a residential laboratory where physicists, computer scientists, chemists, and engineers work with biologists to create synergies



Current Practices: Federal Science Programs



USDA-ARS, DOE-OS, NOAA,
DOD-ONL, NSF, NIH, NASA



Federal R&D Budget – FY 2008

- Total Federal R&D funding = \$142 billion
- Funding for basic research = \$28.4 billion
- Funding for basic research: top six agencies (dollars in millions)
 - HHS/NIH \$15,615
 - NSF 3,993
 - DOE 3,409
 - NASA 2,226
 - DOD 1,428
 - USDA 771

Federal Research Priorities FY 2008

- The American Competitiveness Initiative
 - Two year total new investment of \$2.6 billion for basic research in the physical sciences and engineering (NSF, DoE Office of Science, NIST core)
- Homeland Security
- Energy Security
- Advanced Networking and High-End Computing
- National Nanotechnology Initiative
- Understanding Complex Biological Systems
- Environment
 - Integrated Earth Observations
 - Climate Change Science Program
 - Ocean Action Plan
 - Water Availability and Quality

Hallmarks of Present-day Federal Science Performance Management

- Strategic and annual performance plans well established with linkage to individual performance plans
- Long-term and annual outcome and output measures developed through PART assessments
- Efficiency measures, also through PART
- Emphasis on “leading indicators”, i.e., intermediate outcome indicators
- Emphasis on prospective and retrospective expert review and “community” involvement
- Transparency

Expect**More**.gov

NOAA National Ocean Service

- How to manage knowledge workers in a performance-based system?
- Performance system inherent in doing science is learned in academia (e.g., leadership in community, mentoring students, publications, bringing in funding.) However, need to balance ability to create outputs with ability to affect and influence outcomes – forces organization to think strategically and map it down to the individuals
- Nested measures: Individual outputs contribute to project outcomes that contribute to intermediate outcomes; these work together to have a major strategic outcome for the agency
- NOAA has gone to pay banding (pay for performance) system with standardized output metrics and outcome measures
- Portfolio of measures needs to honor the value system of science. In NOAA's case standardized published metrics were established that would not favor individuals or disciplines; these were publicly vetted and peer-reviewed



Program: National Marine Fisheries Service

Performance Measure: LONG TERM OUTCOME

Number of protected species designated as threatened, endangered, or depleted with stable or increasing population levels.

Explanation

This is a new measure for 2006. The revised performance measure reflects a focus on protected species and the conservation and recovery of protected species through assessments, planning and actions. This measure tracks progress at achieving partial recovery of endangered, threatened or depleted protected species under the jurisdiction of the National Marine Fisheries Service from a baseline of 65 protected species established as of January 1, 2004. Protected species are defined as all marine mammal stocks (except walruses, polar bears, and manatees) and those domestic non-marine mammal species listed as threatened or endangered under the Endangered Species Act that are under the jurisdiction of the National Marine Fisheries Service. Marine Mammal species can be listed as "depleted" under the Marine Mammal Protection Act.

Year	Target	Actual
2004	18	24
2005	20	24
2006	24	25

USDA: Agricultural Research Service

- ARS research is cross-disciplinary and problem focused (2200 PhD scientists)
- 22 National Programs on a five-year planning and performance reporting cycle
 - Stakeholder workshops to define big problems
 - Development of five-year action plan
 - External review process where expert panels review project plans and make recommendations
 - Internal Office of Scientific Quality Review – ongoing quality site reviews
 - After 4.5 years, retrospective assessment of quality and impact (OMB investment criteria)
 - Report feeds into beginning of cycle and workshop
- Use of performance information has led to programmatic changes, i.e., phasing out of bromide program

DOE Office of Science

- 150 PhD scientists mostly manage work that is contracted out to GoCos (national labs) and universities
- DOE appropriators have said ability to articulate goals and measures is a necessary prerequisite to funding. Evaluation is part of program management and in PART review
 - DOE was the guinea pig for PART – 100% of programs two years in a row
 - NOAA and NIH following DOE lead on Leading Indicators Framework
- Good research is *competed, peer-reviewed, and has stakeholder participation*
 - Stakeholder involvement through discussion-based conferences and visiting committees
 - Of total \$3.5 billion budget, only \$40 to \$100 million is earmarked; rest is open to the competitive process
 - All programs subject to rigorous external review – external review is qualitative, albeit highly technical
- Constant effort to ensure doing the best science and constant change in programs, e.g., radiation studies (1986) changed to human genome research and in 2001 changed again to microbial genomics research
 - Flexibility to change and shut down programs since work is done through contractors and universities



U.S. DEPARTMENT OF
ENERGY

Program: Biological and Environmental Research Performance Measure: LONG TERM OUTCOME

Measure: Life Sciences - Provide the fundamental scientific understanding of plants and microbes necessary to develop new robust and transformational basic research strategies for producing biofuels, cleaning up waste, and sequestering carbon. An independent expert panel will conduct a review and rate progress (excellent, good, fair, poor) on a triennial basis.

Explanation: See www.sc.doe.gov/measures for more information.

Year	Target	Actual
2006	Excellent	Excellent
2009	Excellent	
2012	Excellent	
2015	Successful	



Program: Biological and Environmental Research
Performance Measure: ANNUAL OUTPUT

Measure: Determine scalability of laboratory results in field environments
– Determine the dominant processes controlling the fate and transport of contaminants in subsurface environments and develop quantitative numerical models to describe contaminant mobility at the field scale.

Explanation: See www.sc.doe.gov/measures for more information, including a meaningful expansion of the abbreviated nonnumeric targets.

Year	Target	Actual
2002	Sequence	Sequence
2003	Identify	Identify
2004	Modeling	Modeling
2005	Bioremediation test	Bioremediation test
2006	Predictive model	Predictive model
2007	Quantify processes	
2008	ID critical pathways	



Program: Biological and Environmental Research Performance Measure: ANNUAL EFFICIENCY

Measure: Average achieved operation time of the scientific user facilities as a percentage of the total scheduled annual operation time.

Explanation: See www.sc.doe.gov/measures for more information.

Year	Target	Actual
2001	>90%	98%
2002	>90%	97%
2003	>90%	97%
2004	>90%	98%
2005	>90%	100%
2006	>95%	97%
2007	>98%	
2008	>98%	



National Science Foundation
WHERE DISCOVERIES BEGIN

Program: Nanoscale Science and Engineering Performance Measure: OUTCOME

OUTCOME: As qualitatively evaluated by external experts, the successful development of a knowledge base for systematic control of matter at the nanoscale.

Explanation:

The purpose of the program is to support fundamental knowledge creation across disciplinary principles, phenomena, and tools at the nanoscale, and to catalyze science and engineering research and education in emerging areas of nanoscale science and technology. As this research program has to do with long-term basic research in a relatively immature field of science, it is difficult to assess its intellectual results annually or through any quantitative measures. Instead, NSF relies on independent review of relevant experts to monitor whether the research program is appropriately structured and is on track toward the goal of providing an appropriate knowledge base.



Program: Mars Exploration Performance Measure: EFFICIENCY

Cumulative and annual percentage baseline cost overrun on spacecraft under development.

Explanation

NASA's Mars Exploration program conducts scientific exploration of the planet Mars, focusing on the search for water and evidence of past or present life. A key indicator of program efficiency is the degree to which NASA avoids cost overruns on spacecraft under development, since overruns result in cuts or delays to future missions—hence reducing the overall amount of Mars science that can be performed—and/or increase costs to taxpayers. This efficiency measure assesses the degree to which, on average, Mars exploration missions in development will not exceed their baseline costs by more than 5 percent annually or 10 percent cumulatively.



Program: Mars Exploration

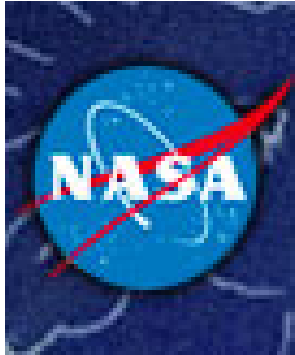
Performance Measure: OUTCOME

Explanation

NASA's Mars Exploration program conducts scientific exploration of the planet Mars, focusing on the search for water and evidence of past or present life. A key indicator of program effectiveness is **NASA's progress in expanding scientific understanding of the planet's evolutionary history and its present-day atmospheric, surface, and interior systems.**

Method

Based on their knowledge of scientific knowledge accrued over a year, external scientific advisors evaluate NASA's Progress annually against this measure (using a green-yellow-red "stoplight" scale), which contributes to the agency's long-term goal of achieving broad scientific understanding of Mars.



Science Community Input

NASA works with the science community to identify questions on the frontiers of science that have profound societal importance, and to which NASA can make a defining contribution.

NASA's Science Mission Directorate acquires community input, independent evaluation, and advice through three principal means. First, the various **boards and committees of the National Research Council** advise NASA on a variety of matters in science, applications, technology, and multi-program planning. Second, the **NASA Advisory Council (NAC)** advises us on program priorities and planning. Finally, **the science community provides input to the Directorate in developing roadmaps** for each science discipline area.

Key Points with Implication for Smithsonian Science



Key Points

- Need strong leadership and internal quality control to provide institutional direction and strategically see relevance of work
 - DOE-OS attributes success in part to its world class scientist administrators such as Ray Orbach, Mildred Dresselhaus and Martha Krebbs
 - Harvard, Berkeley, others have installed internal review committees to provide holistic view and defend institution over division interests
 - USDA-ARS has two-year rotating Quality Officer – seen as a prestigious assignment with duties that include selection of expert panels

Key Points

- Research should be problem-based and interdisciplinary
 - Build capacity around significant problems, i.e., where there are important deficiencies in the world's knowledge base
 - Consider President's list of national priorities

Key Points

- Documented failure is as valuable as documented success: good research requires time discretion, funding discretion, and tolerance for failure
 - Doesn't mean tolerance for incompetence of effort

Key Points

- Success is in competition
 - Of the DOE Office of Science budget (\$3.5B) only \$40 to \$100 million is earmarked; the rest is open to the competitive process, which is very healthy. DOE-OS is constantly re-creating and introspective.
 - Peer pressure at universities and research institutes is described as “so high it would be hard to tolerate if you were unproductive.”
 - Existence of peer pressure depends on the culture of the institution; this culture must be built.

Key Points

- ... But often it is not who you are competing with that is important, but who you are collaborating with
 - Collaborate with federal agencies, universities, etc. – [we are mining the same field for \$\$]
 - Universities, following federal funding changes, are shifting away from individual investigator to large collaborative grants

Key Points

- Expert review by external panels and advisory committees is the best evaluation method for basic science programs
 - The most effective means of evaluating federally funded research is expert review, including quality review, relevance review, and leadership benchmarking (NAS-COSEPUP, 1999)
 - “If you want to determine relevance and validity, you need to bite the bullet and have peer review committees.” (Bostrom)

Key Points

- University model for metrics of scholarship grounded in peer review
 - Dollars brought in is best indicator since that clearly acknowledges work is of the highest caliber and importance
 - Example is U-Penn with 70% HHS funding – it must meet the strong critique of NIH peer reviewed system
 - Publications: consider quality (where published) more than quantity
 - Citation index useful but field-driven; must be mindful of how fields change
 - National and International presence: Sought after speaker? Organizer of keystone symposia?
 - New emphasis on service
 - New emphasis on collaborative projects

Key Points

- Greater stakeholder involvement is a means of survival
 - NASA pulls scientists from around the world to provide input on what areas need to be explored
 - DOE decided five years ago that it could not get any budget increase without a campaign; it worked with universities, labs, and the private sector to demonstrate it is effectively managed and a good investment
 - Museum Conservation Institute (formerly SCMRE) used committee of prestigious materials research experts to prospectively review its draft strategic plan – result was redirection to areas considered by external peer group to be most relevant to the current field

Key Points

- Involve scientists in development of measurement system so that they understand and accept performance measures and see how they can benefit from them
 - At NOAA, standardized published metrics that would not favor individuals or disciplines were publicly vetted and peer-reviewed

Key Points

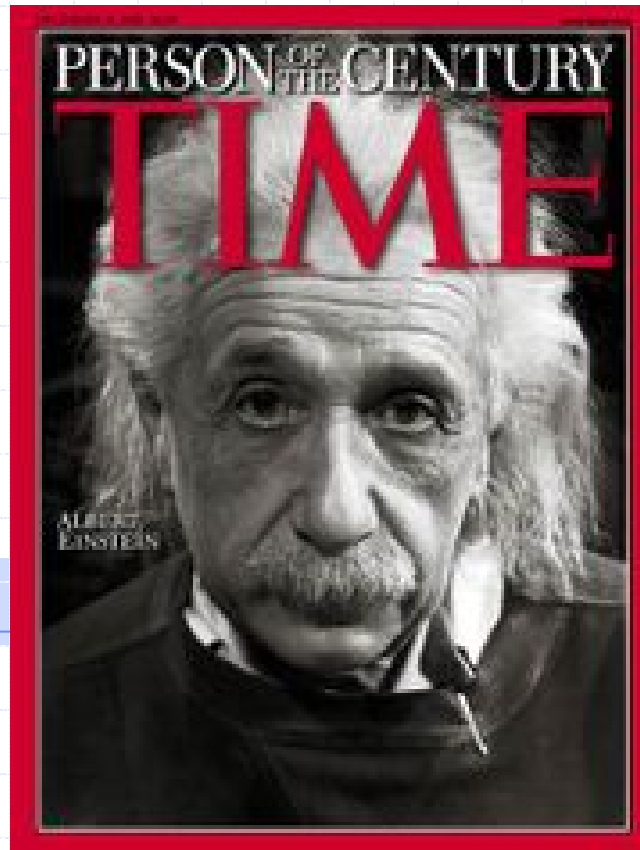
- Utility of logic models
 - Where possible, identify intermediate outcomes in “black box” between individual activities and societal outcomes

Key Points

- Packaging is important
 - Playing the public relations game is necessary if goal is to do more and better science
 - Requires development of efficiency measures (ratio of the outcome or output to inputs of the program) and outcome measures (assessment of results of the program compared to its intended purpose)

Key Points

- And to paraphrase John Donne... No scientist is an island
 - An individual researcher rarely has the level of knowledge to solve significant problems. Ideation and discovery are synergistic
 - Curiosity has to be in an area the institution is interested in... It is done in the social context of collective curiosity
 - Organizations need to develop a track record. If you are working in your own lab and no one comes to see you, you cannot sustain yourself



"Not everything that can be counted counts, and not everything that counts can be counted."

Measuring Science
Presentation to the Smithsonian Science Division, July 12, 2007
Office of Policy and Analysis

Bibliography

- Adams, James and Zvi Griliches. 1996. Measuring science: An exploration. Proceedings of the National Academies of Science. Volume 93, pp. 12664-12670, November 1996
- The American Museum of Natural History. 2002. Policy to Govern the Conditions of Employment, Service, and Responsibilities of the Scientific Staff of The American Museum of Natural History. Approved by the Board of Trustees, 1973, and amended 1998, 2002.
- Branscomb, Lewis M. 1993. U.S. Science and Technology Policy: Issues for the 1990s. <http://www.schwartzman.org.br/simon/scipol/branscomb.pdf>.
- Bush, Vannevar. 1945. Science: The Endless Frontier: A Report to the President on a Program for Postwar Scientific Research. <http://www.jstor.org/view/00228443/ap060035/06a00020/0>.
- Clinton, William J. and Albert Gore, Jr. 1994. Science in the National Interest. August 3, 1994. http://clinton1.nara.gov/White_House/EOP/OSTP/Science/html/Sitni_Home.html. accessed 5/31/2007.
- Cozzens, Susan E. 1999. Results and Responsibility: Science, Society, and GPRA. In American Association for the Advancement of Science: Science and Technology Policy Yearbook. Chapter 16. <http://www.aaas.org/spp/yearbook/chap16.htm>, accessed 5/22/2007.
- _____. 1995. Assessment of Fundamental Science Programs in the Context of the Government Performance and Results Act (GPRA). RAND, Critical Technologies Institute. October 1995.
- Executive Office of the President, Office of Management and Budget. 2007. Program Assessment Rating Tool Guidance 2007-02. Appendix C: Research and Development Program Investment Criteria. January 29, 2007, pp. 72 to 77.
- _____. 2002. The President's Management Agenda, Program Initiative for Better Research and Development Investment Criteria. Fiscal Year 2002, pp. 43-45.
- Executive Office of the President, Office of Science and Technology Policy. 2007. 2008 Federal R&D Budget Highlights.
- _____. 2006. FY 2008 Administration Research and Development Budget Priorities. July 23, 2006

- Feeney, Mary Kathleen and Barry Bozeman. 2005. Taxonomy for science and engineering indicators: a reassessment. *Research Evaluation*. Volume 14, number 3, December 2005
- Feller, Irwin, George Gamota and William Valdez. 2003. Assessing appropriateness: Developing science indicators for basic science offices within mission agencies. *Research Evaluation*, volume 12, number 1, April 2003, pages 71-79. Surrey, England: Beech Tree Publishing.
- Geisler, Eliezer. 2002. R&D Metrics in Technology-Driven Organizations. Paper prepared for the Center for Innovation Management Studies at North Carolina State University. <http://cims.ncsu.edu/documents/rdmetrics.pdf>.
- Godin, Benoit. 2006. Statistics and Science, Technology, and Innovation Policy: How to Get Relevant Indicators. Paper prepared for OECD Blue Sky II Conference, Ottawa, Canada, September 25-27, 2006.
- Godin, Benoit and Christian Dore. 2003. Measuring the Impacts of Science: Beyond the Economic Dimension. http://www.csiic.ca/PDF/Godin_Dore_Impacts.pdf.
- Harvard University, The University Planning Committee for Science and Engineering (UPCSE). 2006. Enhancing Science and Engineering at Harvard. December 2006.
- Howard Hughes Medical Institute. 2006. *Defining Moments*. 2006 Annual Report. <http://www.hhmi.org/catalog/main?action=product&itemId=318>
- Jordan, Gretchen, Sandia National Laboratories. 2005. Contributions of Logic Models to Research, Technology, and Development Policy and Program Evaluation. Presentation at the R&D Evaluation Workshop, Tokyo Japan, June 3, 2005. <http://www.wren-network.net/resources/2005RD.Japan/2005.RD.Japan.htm>
- National Academy of Science, Committee on Science, Engineering, and Public Policy (COSEPUP). 2000. *Experiments in International Benchmarking of U.S. Research Fields*. Washington, DC: National Academy Press. <http://www.nap.edu/catalog/9784.html>
- _____. 1999. *Evaluating Federal Research Programs: Research and the Government Performance and Results Act*. Washington, DC: National Academy Press. <http://www.nap.edu/catalog/6416.html>
- _____. 1993. *Science, Technology, and the Federal Government: National Goals for a New Era*. Washington, DC: National Academy Press. <http://www.nap.edu/catalog/9481.html>
- National Research Council, Committee on Metrics for Global Change Research, Climate Research Committee. 2005. Thinking Strategically: The Appropriate Use of Metrics for the Climate Change Science Program. <http://www.nap.edu/catalog/11292.html>

- National Science Board. 2006. *Science and Engineering Indicators 2006*. Two volumes. Arlington, VA: National Science Foundation (volume 1, NSB 06-01; volume 2, NSB 06-01A).
- National Science Foundation. 2007. FY 2006 Performance Highlights. <http://www.nsf.gov/pubs/2007/nsf0711/nsf0711.jsp>
- National Science and Technology Council. 1996. *Assessing Fundamental Science: A Report from the Subcommittee on Research, Committee on Fundamental Science*. July 1996. <http://www.nsf.gov/statistics/ostp/assess/> accessed 5/24/2007.
- Popper, Steven W. 1995. *Economic Approaches to Measuring the Performance and Benefits of Fundamental Science*. RAND, Critical Technologies Institute. October 1995.
- Scott, Alister. 2006. *Peer Review and the Relevance of Science*. University of Sussex, SPRU – Science & Technology Policy Research. Paper No. 145, February 2006.
- Tassey, Gregory. 2003. *Methods for Assessing the Economic Impacts of Government R&D*. National Institute of Standards and Technology. <http://www.wren-network.net/resources/MethodsforAssessingtheEconomicImpactsofGovernmentR-D.pdf>
- Valdez, William. 2005. *DOE-SC Evaluation, Research, and Policy Development*. Presentation given at Korea Institute of S&T Evaluation and Planning (KISTEP) – Washington Research Evaluation Network (WREN) Workshop/International Symposium, May 30-31, 2005. <http://www.wren-network.net/resources/2005kistep.htm> accessed 6/21/2007.
- Valdez, William. 2005. *Evaluation of Public Sector R&D in the United States: Lessons Learned from GPRA and the Program Assessment Rating Tool*. Presentation at the R&D Evaluation Workshop, Tokyo Japan, June 3, 2005. <http://www.wren-network.net/resources/2005RD.Japan/2005.RD.Japan.htm>
- Wagner, Caroline S. 1995. *Techniques and Methods for Assessing the International Standing of U.S. Science*. RAND, Critical Technologies Institute. October 1995.