Frontiers of evaluation: United States*

Susan E. Cozzens
Ivan Allen College of Liberal Arts, Georgia Institute of Technology, USA

Recebido: 22/03/2011 Versão revisada (entregue): 02/09/2011 Aprovado: 07/12/2011

Abstract

For someone from the United States, addressing the theme of the "frontiers of evaluation" is a pleasure. Frontiers are an important part of the American self-concept. We have always seen ourselves moving west, into new territory that is expected to bring freedom and prosperity. This paper will trace that moving frontier in research evaluation in the United States, starting with a little history, moving towards the present, and peeking into the future. It begins with the transition from program evaluations to accountability systems, then turns to the current effort to establish a Science of Science Policy that will provide the research base for future evaluation techniques. I will illustrate current practice with several examples: complex rating systems, logic models, economic estimations, and mapping, before describing a new dataset still in preparation. I will conclude with some observations about where U.S. research evaluation has been and is going.

Keywords | Evaluation; Metrics; Logic models; Accountability.

JEL Codes | H5.

* The author appreciates the helpful comments of anonymous referees.

Fronteiras da avaliação: Estados Unidos

RESUMO

Para um cidadão dos Estados Unidos, é um prazer falar sobre o tema das "fronteiras da avaliação". Fronteiras constituem uma parte importante da autoimagem dos americanos. Sempre tivemos o anseio de marchar para o oeste e de conquistar territórios novos que trarão liberdade e prosperidade. Este trabalho traça a fronteira móvel da avaliação da pesquisa nos Estados Unidos, começando com um pouco de história, chegando até o presente e vislumbrando o futuro. Inicialmente discute-se a transição de avaliações de programas para sistemas de prestação de contas e transparência (*accountability*), para em seguida enfocar o esforço atual de estabelecimento de uma Política de Ciência da Ciência que irá deitar as bases para as futuras técnicas de avaliação no campo da pesquisa. Ilustra-se a prática atual com vários exemplos: sistemas complexos de classificação, modelos de lógica, estimativas econômicas e mapeamento. Em seguida, é descrito um novo conjunto de dados que ainda está sendo desenvolvido. Nas conclusões, são apresentadas algumas observações sobre a história e o futuro da avaliação da pesquisa nos Estados Unidos.

PALAVRAS-CHAVE | Avaliação; Métrica; Modelos de lógica; Prestação de contas (*accountability*).

CÓDIGOS JEL | H5.

1. Introdução

1. History of U.S. research evaluation

Federal support for research in the United States dates back to the late 1940s and early 1950s, with the formation of the Office of Naval Research, the National Institutes of Health, and the National Science Foundation (ENGLAND, 1982; HARDEN, 1986; SAPOLSKY, 1990). Program evaluation started soon afterwards, with the earliest surviving example being the program reviews carried out for the National Institute of Standards and Technology (NIST) by a standing board of the National Academies of Science (NAS); these date from the mid-1950s. The auspices of the National Academies certified the expertise of this process, which used external panels of experts to produce qualitative assessments that were delivered to NIST management (COZZENS, 1997). Other agencies started program evaluation with internal processes, as was the case with NASA (the National Aeronautics and Space Administration), which did its own annual program reviews with presentations to management (COZZENS, 1987).

As federal investment grew, the demand for more systematic information on its effectiveness grew, too. Public officials needed methods that could be widely understood to justify spending taxpayer dollars. The methods community responded in the 1960s. One of the earliest attempts to develop a method was project Hindsight, a Defense Department study that traced the factors that contributed to several prominent examples of new defense technologies (SHERWIN, 1967). Since this study showed that it was mostly technology that contributed to technology, over the relatively short space of several decades, research agencies worried that all the funding would be redirected to applied development efforts. In a study called TRACES (Technology in Retrospect and Critical Events in Science), the National Science Foundation (NSF) therefore used a similar methodology but focusing on civilian technologies and tracking contributions over a longer time period – and accordingly demonstrated that basic research was an essential element even of technological advances (IITRI,1968). A similar study by Comroe and Dripps (1976) demonstrated the importance of basic research in the biomedical area .

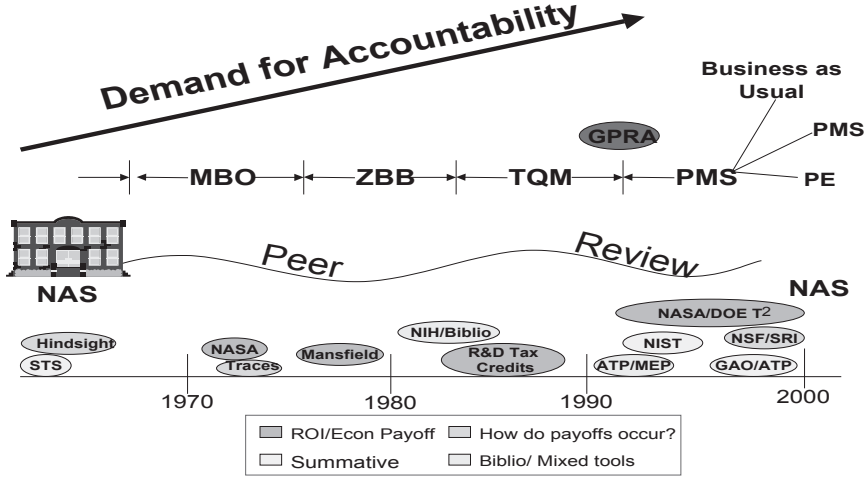
By the 1970s, with the advent of large-scale bibliographic databases including citations from one paper to another, the National Institutes of Health (NIH) and NSF began to build literature-based ("bibliometric") datasets for use in evaluation at program, institute, and national levels. Early validation work demonstrated that

citations were a reasonably good proxy measure of scientific quality, and the era of publication and citations counts as primary evaluation data was born (NARIN, 1976). NIH used the data in a series of reports on publications associated with the various institutes, and NSF used them in the National Science Board's Science Indicators series, as well as in specialized studies such as the original converging indicators study, of the Materials Research Centers (Ling and et-al. 1978).

Figure 1, from a paper by J. David Roessner (2002), depicts the development of research evaluation methods in the United States through about 2000. The demand for accountability rises steadily, as it has since that time. A line of initials marches across the upper part of the figure, naming a series of performance management efforts in the U.S. federal government: Management by Objectives (MBO), Zero Based Budgeting (ZBB), Total Quality Management (TQM), and other performance management systems, finally splitting the traditions into program evaluation and performance monitoring. Constant throughout the period are peer reviews done by the National Academy of Sciences – not just the series already mentioned, but also peer assessments of individual programs. Finally, at the bottom of the figure, we see the evolution of systematic methods: Hindsight, TRACES, and the NIH bibliometric data to a set of economic studies focused on more industry-oriented programs (discussed further below).

FIGURE 1

Milestones in the history of research evaluation in the U. S.



Source: Roessner, Outcome Measurement in the United States, 2002. Reprinted with permission of the publisher.

2. GPRA to PART to SOSP: the march of acronyms

A major milestone in U.S. research evaluation, not noted in Roessner's figure, slipped quietly onto the scene in 1993. This was the Government Performance and Results Act (GPRA) (COZZENS, 1999). In contrast to the previous performance-based management efforts such as MBO and ZBB, GPRA was a law, not just an executive order. It mandated every federal agency to do a strategic plan every three years, a performance plan annually, and a performance report every year as well. Agencies were required to set quantitative performance indicators and set annual targets for performance.

This regime was very far away from the evaluation practices of agencies in 1993.[1] Many agencies were doing program reviews, either external as in the case of the National Institute of Standards and Technology (NIST) or internal, in Defense and USDA (the U.S. Department of Agriculture). Under pressure from Congress to demonstrate the quality of its research, the Department of Energy (DOE) had developed the most systematized set of reviews, using external panels and a carefully designed rating form that turned qualitative expert judgments into standardized scores on established criteria (COZZENS, 1987). The basic research agencies, NSF and NIH, resisted the idea that they could do such regular reviews of performance of their programs on the ground that one could not predict where they would have their impacts, but they did sponsor occasional specialized studies. In all these cases, programs were evaluated against their own goals (no comparison groups), on criteria that included quality, relevance, and "productivity," generally understood as quantity of appropriate outputs such as reports or publications. The expert panel methods relied on basic program information, while the specialized studies used more sophisticated techniques such as normalized citation counts, journal categories, and clustering and mapping using citations (SMALL, 1973).

The research agencies, and the research branches of larger mission agencies, were thus ill-equipped in 1993 to address the requirements of GPRA. Strategic planning was anathema at the research agencies. NSF's Director had recently received approval from the National Science Board, its governing agency, for a very general plan – but only with difficulty. An NIH strategic plan developed by the recently-departed director had been scuttled as soon as she left; within days, it was

1 My students and I had just completed a set of interviews in agencies on their evaluation methods when the Act was passed. This description is based on those interviews.

impossible to locate a copy on the desks of NIH officials. Because GPRA was linked to the budget process, and implemented by the Office of Management and Budget (OMB) in the White House, budget staff members were in charge of implementation; but they almost never knew anything about the methods for assessing performance of research programs. Measures were available, but not well suited to setting annual performance goals (COZZENS, 1997). The agencies tried to protest that research should be exempted from the law, but their protests met deaf ears and they settled in to figure out how to comply – and perhaps even benefit from it.

During the pilot phase of GPRA implementation, the evaluation offices of the various agencies formed a self-help network, and after some discussion adopted a framework for performance measurement based on the experience of the Army Research Laboratory.[2] They also studied closely a pilot project at NSF with use of the "alternative format" for performance reporting included in a footnote of GRPA. The alternative format allowed for more qualitative goal setting. The social capital built with the early inter-agency GPRA network was systematized a decade later in the Washington Research Evaluation Network (WREN), which held regular meetings to allow agencies to share best practices.[3]

GPRA, in the meantime, was still on the books. But every new administration likes to develop its own version of performance management, and the Bush administration indeed buried GPRA under a new process called PART, the Program Assessment Rating Tool. PART provided a set of criteria for rating programs, and led to infamous "green," "yellow," and "red" ratings of agency progress towards goals. The research agencies again protested that the generic criteria did not apply to them, and OMB responded by developing a specialized set of criteria, again focused on quality and relevant, but this time including "performance," by which they meant progress towards specified outcomes. These criteria led to a variety of discussions with agencies, including some pressure on the Environmental Protection Agency to count project completion as "performance" (COMMITTEE ON EVALUATING THE EFFICIENCY OF RESEARCH AND DEVELOPMENT PROGRAMS AT THE U.S. ENVIRONMENTAL PROTECTION AGENCY, 2008) and narrowing performance reporting for NIH to a few specific programs where concrete objectives could be met in relatively short periods of time.

2 See <http://govinfo.library.unt.edu/npr/library/studies/casearla.pdf>, accessed December 30, 2010.

3 <http://www.wren-network.net/>.

Late in the Bush Administration there was a development in the White House science office that might create a quantum leap for U.S. research evaluation. The President's science advisor, Dr. Jack Marburger (2005), expressed the need for a better base of empirical evidence to support his recommendations on science funding and instruments. His call led to the Science of Science Policy (SOSP) initiative, which built a network based on WREN but was also accompanied by a substantial program of funding at NSF for basic research on the science policy dynamics. (The program is called SciSIP – the Science of Science and Innovation Policy.) I will return to the influence of this development in the final section of the paper.

3. Current practices

Through all this evolution and network-building, what has changed in the practice of research evaluation at the federal level in the United States? Not much. The bread-and-butter method is still evaluation by expert panel, with variable inputs. Mostly, these panels are still working with program-generated information on inputs (funding, people, project names, etc.) plus a bit of output information (papers and patents). The research agencies (NSF and NIH) are still not using strategic planning much as a tool, and at first clung to using anecdotal information on program accomplishments as performance reporting ("stories of accomplishment" at NIH, "golden nuggets" at NSF), with some ambivalence. Strategic planning is going on in all the mission agencies, but since many of these spend much of their money in government laboratories, and since government laboratory personnel are very hard to re-orient, the alignment between actual research activities and strategic plans is usually far from perfect. Nonetheless, some movement towards more sophisticated forms of performance assessment is emerging.

3.1. Complex rating systems

One example is the elaboration of expert rating systems. The Sea Grant Program at the National Oceanic and Atmospheric Administration illustrates this (COMMITTEE ON THE EVALUATION OF THE SEA GRANT PROGRAM REVIEW PROCESS, 2006). Sea Grant is a program of block funding to states, that is, the federal government gives a lump sum of money to the state with expectations for delivering certain kinds of research activities to protect coasts and

fisheries. Because of its connection to states, and the states' connections to elected members of Congress, the program is under continual political pressure – often a factor in the development of evaluation processes. Indeed, Sea Grant's national board has developed a highly complex system of rating by external experts, with quantified results that can be used for performance funding. Board 1 shows the areas in which these panels rate performance for each state Sea Grant program on a three-year evaluation cycle.

BOARD 1

Sea grant program review areas

Organizing and Managing the Program (20%)	Leadership of the Program (6%) Institutional Setting and Support (4%) Project Selection (2%) Recruiting Talent (3%) Effective and Integrated Program Components (5%)
Connecting Sea Grant with Users (20%)	Engagement with Appropriate User Communities (15%) Partnerships (5%)
• Effective & Aggressive Long-Range Planning (10%)	Strategic Planning Process (4%) Strategic Plan Quality (4%) Implementation Plan (2%)
• Producing Significant Results (50%)	Contributions to Science and Technology (10%) Contributions to Extension, Communications and Education (10%) Impact on Society, the Economy, and the Environment (25%) Success in Achieving Planned Program Outcomes (5%)

The full description for each area in Board 1 includes qualitative "benchmarks," as illustrated in Board 2.

The Sea Grant process attempts to achieve standardization of the use of the rating scheme by assigning chairs from the national board and an 18-page manual that spells out the benchmarks and provides standard descriptions of performance levels (these appear in Board 3). Despite all the care taken in its design, the system was highly controversial because of its link to the distribution of resources – which the states would rather have distributed on a basis that is independent of performance, regardless of how competently it is judged.

BOARD 2

Illustration: criterion and benchmarks

Institutional Setting and Support The emphasis for this criterion should be placed both on the effectiveness of the reporting relationship for the Sea Grant program within the institution and on the overall level of support provided by the institution. In general, though, the expectation is that the program reports to the highest possible level within the institution.
• **Expected Performance Benchmark** – The program is located at a high enough level within the university to enable it to operate effectively within the institution and externally with all sponsors, partners, and constituents. – The institution provides the support necessary for the Sea Grant program to operate efficiently as a statewide program.
• **Indicators of Performance** – Setting of the program within the university or consortium organization and reporting structure – Program infrastructure (space, equipment, available resources)

BOARD 3

Ratings levels in the sea grant evaluation system

Needs Improvement – In general, performance does not reach the benchmark for this sub-element. The [program assessment team] will identify specific problem areas that need to be addressed.
Meets Benchmark – In general, performance meets, but does not exceed, the benchmark for this sub-element.
Exceeds Benchmark – In general, performance goes beyond what would be required to simply meet the benchmark for this sub-element.
Highest Performance – Performance goes well beyond the benchmark for this sub-element and is outstanding in all areas.

3.2. Logic models

Another planning and evaluation tool that U.S. research agencies have begun to use is the logic model. This is a long-standing approach in program evaluation generally (see MCLAUGHLIN; JORDAN, 1999), but it has only begun to be used by

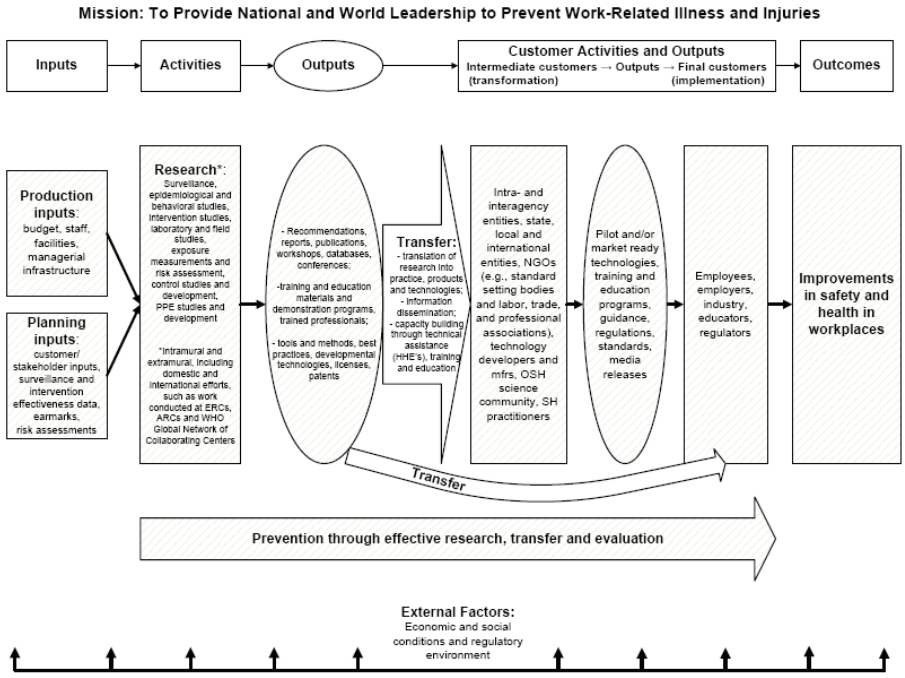
research agencies in the last decade or so. An illustration comes from the National Institute of Occupational Safety and Health (NIOSH). Again under pressure from outside (this time from the Bush administration's Office of Management and Budget), NIOSH undertook an ambitious set of external reviews of its programs, carried out by the Institute of Medicine (part of the National Academies complex). What was particularly praiseworthy in the NIOSH reviews was the request for the review panels to hold its programs accountable for outcomes, not just outputs. That is, NIOSH took genuine accountability for improving the health and safety of U.S. workers. But to do so, it needed to specify how its research programs were actually expected to achieve that goal. The laudable effort was further impeded by the lack of good measures of the hoped-for health and safety outcomes.

The solution NIOSH adopted was to develop a set of logic models for its programs, complete with intermediate outcomes.[4] (These can be seen for the NIOSH example in Figure 2.) A logic model is a linear representation of how a program is supposed to work. It begins at the left with input and moves through activities and outputs (the immediate, tangible results produced by the program). At the far right are the outcomes the program expects – what it promises to deliver to the public over the longer term. In between are the intermediate outcomes – steps along plausible pathways from outputs to outcomes. The advantage of intermediate outcomes in program evaluation is that they can be observed within the time frame of the evaluation, usually three to five years retroactively. In the NIOSH model, they included use of research results by regulatory agencies and employers in worksite health programs.

Logic models and intermediate outcome measures are intimately related. It is the linearity of the logic model that allows for specification of intermediate outcomes. But linear models of how research creates benefits have long been considered outdated by scholars of the research and innovation process. The intermediate outcome measures that NIOSH used focused attention on short-term, linear effects and drew attention away from less linear and longer terms ones, such as the benefits of training a workforce for occupational safety and health research, strengthening the network of researchers and practitioners concerned with these issues, and changing frameworks for thinking about the hazards. Since these kinds of effects could not be measured in the framework used for the NIOSH reviews, the reports produced no information, positive or negative, on NIOSH performance in these dimensions.

4 NIOSH hired an outside consultant firm to help it develop its generic logic model and to facilitate individual programs in developing their own.

FIGURE 2
The NIOSH logic model



Source: Reprinted with permission from *Evaluating Occupational Health and Safety Research Programs: Framework and Next Steps, 2009,* by the National Academy of Sciences, Courtesy of National Academies Press, Washington, D.C.

3.3. Economic analysis

While agencies with research missions tied to public goals like occupational safety and health or sustainable coastal ecologies have been experimenting with expert-based assessment models and systems, U.S. federal programs more closely tied to the work of private, technology-based firms in the economy have been developing methods for estimating their economic returns. The lead agency for many years in this regard was the National Institute of Standards and Technology (NIST), the home of the Advanced Technology Program (ATP), the government's lead civilian program for stimulating development of new products and processes. The ATP evaluation staff funded many studies, using a variety of methods, to demonstrate the impacts of the program – all with significant, positive results.

Figure 3 gives an example of the logic the ATP studies used to track benefits of its projects, not just for the firms that received ATP funding, but also for other firms and for the public (the "spillover" effects). Measuring the spillover effects was critical to demonstrating the rationale for public investments in these projects, which ultimately took shape as privately-owned products and processes. How did ATP measure these? Even its primer on evaluation techniques[5] does not explain the techniques in detail, but instead provides a list of references:

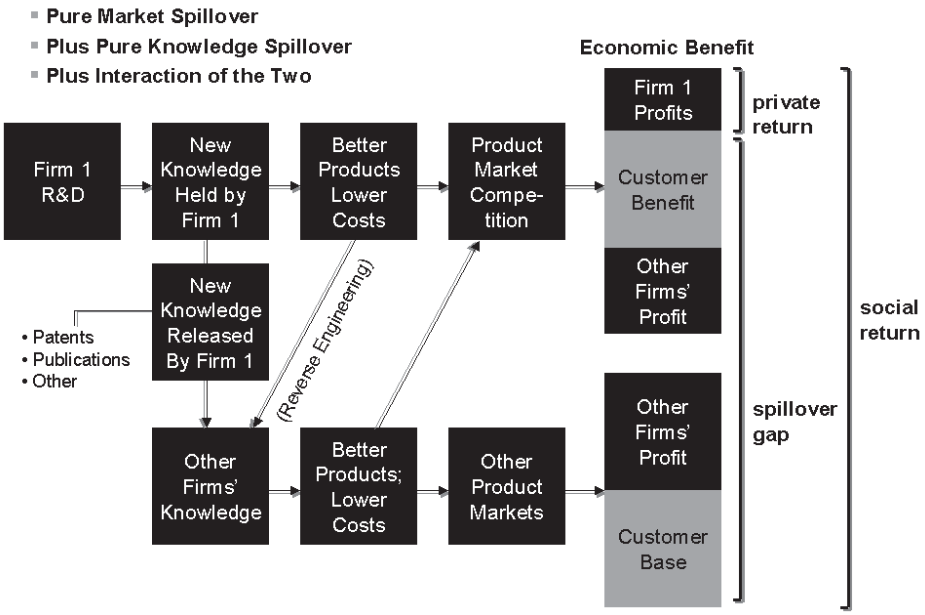
- Starting point: (MANSFIELD et al., 1977) approach
- Cost index method (AUSTIN; MACAULEY, 2000)
  - estimates market spillovers to consumers
  - avoids the requirement for market data in Mansfield approach.
- R&D Economic Value of Knowledge Spillovers (DENG, 2005)
  - Uses patent citations at the firm level, building on work by Jaffe and Lerner (2001) and others.
- Combination (MANSFIELD, 1996)
  - Combining market and knowledge spillover assessment in program or project case studies
  - Single benefit-cost framework
  - More comprehensive results

The implication is that if one wants to measure economic effects, one needs to hire an economist – and so ATP did, in large numbers. Over time, it built up a portfolio of studies using both case study and larger econometric techniques to demonstrate the effectiveness of the program. Unfortunately, the program fell prey to political opposition despite these solid results.

5 Ruegg, R. and I. Feller (2003). A Toolkit for Evaluating Public R&D Investment: Models, Methods, and Findings from ATP's First Decade. Washington, D.C., National Institute of Standards and Technology. <http://www.atp.nist.gov/eao/gcr03-857/contents.htm.>.

FIGURE 3

Logic Model for Spillover Effects of Advanced Technology Projects



Source: Ruegg and Feller (2003). Reprinted with permission.

3.4. Mapping

Although techniques for making spatial representations of the intellectual content of science have been in existence since the 1970s, the new attention to the "science" of science policy has also drawn new attention to this capability. Indeed, information processing capacities have changed enough to label the current versions as a new generation of science mapping. NSF has been particularly interested in exploring the intersection of visualization studies in computer science with traditional literature-based databases and emerging web-based datasets (LANE; COZZENS, 2008). This area is truly a frontier of research evaluation in the United States in several senses. First, new analysts are entering from computer science (sometimes lacking the models and substantive understanding of the realities and dynamics of research and the limitations of the underlying datasets). Second, the actual applications of the maps for practical problems of government planning, funding, or implementing research programs are still under exploration. But isn't that what a frontier is about?

An example of a recent attempt to apply these techniques comes from the Environmental Protection Agency (EPA). EPA has been doing traditional expert review of its programs for some time, with only limited experimentation with new measures of the results of those programs. Recently, however, it decided to try mapping in connection with an evaluation of its Drinking Water Research Program. This program employs about 170 researchers, with an annual budget of about 47 million dollars. The staff working on the evaluation decided to try science mapping to answer the following questions:

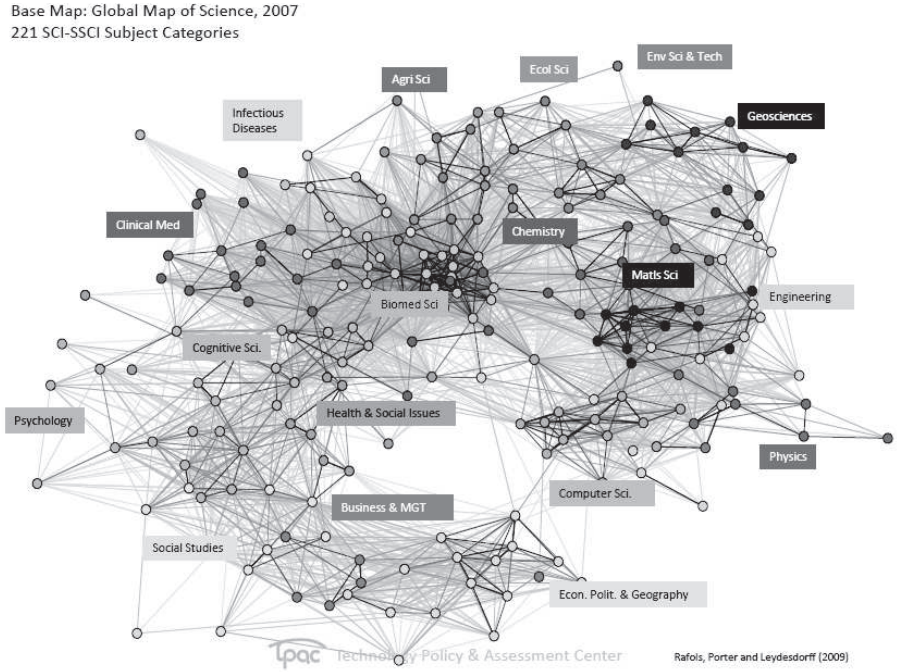
- What is the current discipline make up within the research program?
- How has this changed over time?
  - Map of research program's discipline expertise as represented by publication outlets.
- How can this information be used to
  - Retrospectively evaluate program performance?
  - Prospectively evaluate research capability?
  - Identify workforce planning opportunities?

They began with a "base map" of science produced by other scholars (Rafols, Porter et al. 2010). Against this base map, they marked the journals where publications from the Drinking Water Program published their results (see Figure Four).

An initial introduction of this map to the EPA committee that designs evaluations for the agency illustrated several of the challenges that science mapping has encountered over its decades of existence. First, the meaning of the map is not clear without extensive explanation of the procedure followed, some of which is quite complicated. Some observers remained skeptical as long as they did not fully understand what the circles and lines came from and what they meant – a task that takes a long lecture, not a short, policy-oriented briefing. Second, some members of the panel questioned whether journal disciplines – the main data about EPA researchers revealed on the map – were good indicators of the disciplines of researchers. Finally, it remained unclear to the committee what the discipline map might say about the program's effectiveness. In short, while the picture was very pretty, it was not clear what it said in the evaluation context on the ground.

FIGURE 4

EPA Drinking Water Program Discipline Map



Source: Levine and Miller (2010). Drinking water related research: USEPA 1979-2010. Unpublished raw data. Reproduced with permission.

4. Beyond the frontier

4.1. STAR Metrics

Beyond the frontier of evaluation for research programs in the United States lie the tools that the Science of Science Policy initiative is trying to develop, for the future of research evaluation. Prominent among this at the current time is the STAR Metrics project[6] (**S**cience and **T**echnology for **A**merica's **R**einvestment: **M**easuring the **E**ffec**T**s of **R**esearch on **I**nnovation, **C**ompetitiveness and **S**cience), a university partnership to document the outcomes of science investments to the public.[7] The project is currently in Phase I, focusing on developing uniform, auditable and standardized measures of the initial impact of stimulus funding on science spending on

6 Much of this text comes directly from slides prepared by the program to explain its structure and aims.

7 See <http://www.nih.gov/news/health/jun2010/od-01.htm>, accessed December 30, 2010.

job creation. It is minimizing costs by downloading data directly from university administrative systems.

In Phase II, the project will move on to collaborative development of measures of the impact of federal science investment on:

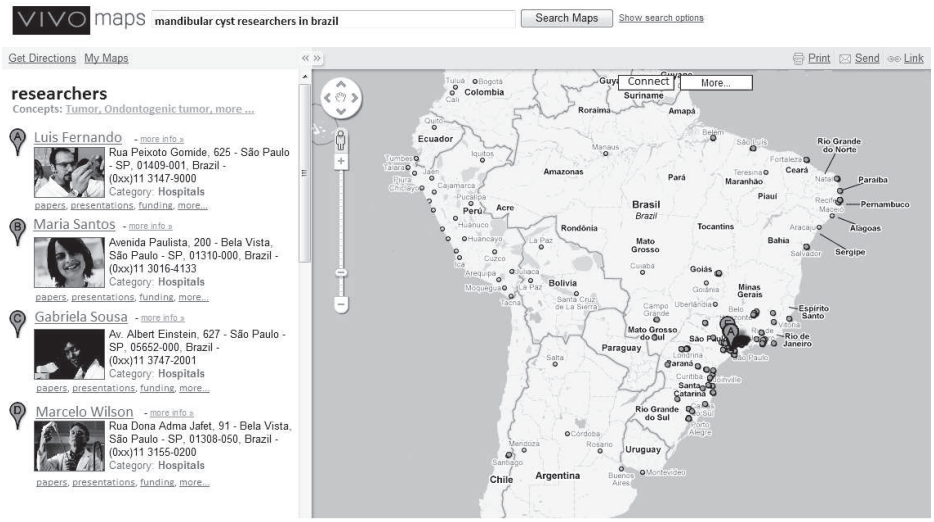
- economic growth (through patents, firm start ups and other measures),
- workforce outcomes (through student mobility and employment),
- scientific knowledge (such as publications and citations) and, later,
- social outcomes (such as health and environment).

The basic approach is to build on existing investments, including the Brazilian Lattes database, which serves as an example of what can be done; Vivo, software for national networking;[8] and ideas offered by the STAR Metrics partner institutions. Those institutions are working collaboratively to identify the best approach and address privacy and technical issues and to plan dissemination and links to other federal activities. An inter-agency working group is leading the effort.

In the future, the working group intends STAR Metrics to serve as a large-scale retrieval and networking tool (see Figure 5). The data could be used to create biosketches, curricula vitae, annual reports, and department and research group web sites, or to populate profiles in collaborative tools – portals, wikis, etc.

FIGURE 5

Finding Global Science Connections through STAR Metrics



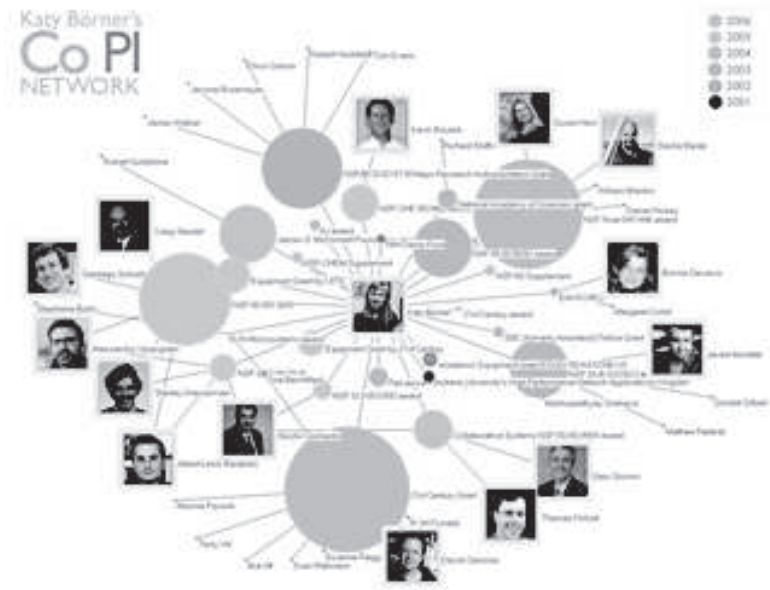
Source: Reproduced by permssion of Mike Conlon, VIVO team lead (http://vivo.ufl.edu/)

8 <http://vivo.cornell.edu>.

The kind of data the system will capture could eventually lead to new power in analyzing social relations in science, as Figure 6 illustrates.

FIGURE 6

A sample PI/Co-PI network



Source: Katy Borner's Co-PI Network 2001-2006. Courtesy of Cyberinfrastructure for Network Science Center (http://cns.iu.edu). More recent collaborations are lighter in color. Larger circles indicate larger awards.

4.2. Measuring outcomes

The SOSP initiative, with its partner the SciSIP funding program at NSF, is actively pushing beyond the frontier the ability to measure the results of research funding in the United States – that is, the core of evaluation information. A recent workshop exhibited the state of the art for Washington audiences.[9] Review papers covered four areas: economic benefits, technology development and employment, S&T workforce development, and social, health, and environment benefits. The topics clearly appeared in descending order of maturity.

In the first area, economists reviewed the decades of measurement of returns from agricultural research; new methods of estimating indirect effects through job

9 <http://www.nsf.gov/sbe/sosp/>, retrieved November 2010.

creation and increases in productivity; the problems in estimating the payoff to R&D with production function/growth accounting, offering some solutions to the problems based on ongoing work with the NSF's Survey of Industrial Research and Development (SIRD); and frontier tools and applications in measuring the impact of science policy on the rate and direction of cumulative research. Others previewed evaluation frameworks for examining new policy instruments like prizes; reviewed new data sources for large-scale analysis of science policy outcomes; and evaluated the state of data on personnel, including how STAR Metrics might contribute to its depth and quality. Finally, a set of papers addressed evaluation tools for social, environmental, and health benefits from research, succeeding largely in illustrating the dearth of data and techniques for this area. Even the presenter on agricultural research acknowledged limited policy attention to the results of this long-standing area with extremely rich data. The frontier is still very much open, both in methods and in policy application and use.

5. Conclusions

The more we look at present efforts in the United States, the larger the gap becomes between theories of science and technology dynamics, quantitative forms of analysis, and the realities of research program evaluation in the agencies. Since the bread and butter method for evaluation is still an expert panel, researchers from the science and engineering community are considered essential sources of information for evaluation. But they are not trained in the analytic methods being developed. The evaluation task will never be handed over completely to outside analysts with sophisticated mapping and analytic techniques, however. The work of agency evaluation staff thus centers on integrating complicated analytic methods with expertise in the field being evaluated. This is likely to become a bigger and bigger part of their responsibilities.

A community of practice – that is, a network of interaction that involves both practitioners and academic specialists in a dialogue about what is both valid and useful -- can help evaluation staff develop those skills. With an influx of new resources, a community of this sort is being built through the work of the Science of Science Policy Subcommittee. The frontier of research evaluation in the United States is thus expanding and moving rapidly forward. Times are exciting in Washington.

Bibliographical references

AUSTIN, D.; MACAULEY, M. E*stimating future consumer benefits from ATP-funded innovation*: the case of digital data storage. Gaithersburg, MD, National Institute of Standards and Technology, 2000.

COMMITTEE ON EVALUATING THE EFFICIENCY OF RESEARCH AND DEVELOPMENT PROGRAMS AT THE U.S. ENVIRONMENTAL PROTECTION AGENCY, N. R. C. *Evaluating research efficiency in the U.S. environmental protection agency*. Washington, DC, National Academies Press, 2008.

COMMITTEE ON THE EVALUATION OF THE SEA GRANT PROGRAM REVIEW PROCESS, N. R. C. *Evaluation of the Sea Grant Program Review Process*. Washington, D.C.: National Academies Press, 2006.

COMROE, J. H.; DRIPPS, R. D. Scientific basis for support of biomedical science. *Science*, v. 192, n. 4.235, p. 105-111, 1976.

COZZENS, S. E. Expert review in evaluating programs. *Science and Public Policy*, v. 14, n. 2, p. 71-81, 1987.

_____. The Knowledge pool: measurement challenges in evaluating fundamental research programs. *Evaluation and Program Planning*, v. 20, n. 1, p. 77-89, 1997.

_____. Results and responsibility: science, society, and GPRA. In: A. TEICH, H.; NELSON, S. *Science and technology policy yearbook, 1998*.. Washington, D.C.: American Association for the Advancement of Science, 1999.

DENG, Y. The Value of knowledge flows: evidence from patent citations data. *Computing in Economics and Finance 2005*. Society for Computational Economics, 2005.

ENGLAND, J. M. *A patron for pure science:* the National Science Foundation's formative years, 1945-57. Washington, D.C.: National Science Foundation, 1982.

HARDEN, V. A. *Inventing the NIH*: federal biomedical research policy, 1887-1937. Baltimore, Md.: Johns Hopkins University Press, 1986.

IITRI – Illinois Institute of Technology Research Institute. *Technology in retrospect and critical events in science*. Washington, D.C.: National Science Foundation, 1968.

JAFFE, A. B.; LERNER, J. Reinventing public R&D: patent policy and the commercialization of national laboratory technologies. *Rand Journal of Economics*, v. 32, n. 1, p. 167-198, 2001.

LANE, J.; COZZENS, S. *A deeper look at the visualization of scientific discovery in the federal context by NSF*. Washington, D.C.: National Science Foundation, 2008. Available in: <http://scienceofsciencepolicy.net/media/p/188.aspx>.

LING, J. T. et al. *Evaluative study of the materials research laboratory program*. McLean, Va., The MITRE Corporation, 1978 (Summary report).

MANSFIELD, E. *Estimating social and private returns from innovations based on the advanced technology program*: problems and opportunities. Gaithersburg, MD: National Institute of Standards and Technology, 1996.

MANSFIELD, E. et al. Social and private rates of return from industrial innovations. *Quarterly Journal of Economics*, v. 91, n. 2, p. 221-240, 1977.

MARBURGER, J. Speech to AAAS Forum on Science and Technology Policy. Washington, D.C., 2005. Available in: <http://www.scienceofsciencepolicy.net/media/p/59.aspx>.

MCLAUGHLIN, J. A.; JORDAN, G. B. Logic models: a tool for telling your program's performance story. *Evaluation and Program Planning*, v. 22, n. 1, p. 65-72, 1999.

NARIN, F. *Evaluative bibliometrics*: the use of publication and citation analysis in the evaluation of scientific activity. Cherry Hill, New Jersey: Computer Horizons, Inc., 1976.

RAFOLS, I. et al. Science overlay maps: a new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, v. 61, n. 9, p. 1.871-1.887, 2010.

ROESSNER, J. D. Outcome measurement in the USA: state of the art. *Research Evaluation*, v. 11, n. 2, p. 85-93, 2002.

RUEGG, R.;FELLER, I. *A toolkit for evaluating public R&D investment*: models, methods, and findings from ATP's first decade. Washington, D.C.: National Institute of Standards and Technology, 2003. Available in: <http://www.atp.nist.gov/eao/gcr03-857/contents.htm>.

SAPOLSKY, H. M. *Science and the navy*: the history of the Office of Naval Research. Princeton, N.J.: Princeton University Press, 1990.

SHERWIN, C. W.; ISENSON, R. S. Project Hindsight: Defense Department Study of the Utility of Research. *Science*, n. 156, p. 1.571-1.577, 1967.

SMALL, H. Cocitation in scientific literature – new measure of relationship between 2 documents. *Journal of the American Society for Information Science*, v. 24, n. 4, p. 265-269, 1973.

CORRESPONDENCE ADDRESS:
Susan E. Cozzens – scozzens@gatech.edu
School of Public Policy, Georgia Institute of Technology
30332-0345 – 685 Cherry Street, Atlanta, GA