

# 32 Evaluating Federal Research Programs: Research and the Government Performance and Results Act

**Committee on Science, Engineering, and Public  
Policy, National Academy of Sciences/National  
Academy of Engineering/Institute of Medicine**

## Executive Summary

The Government Performance and Results Act (GPRA), enacted in 1993, focuses agency and oversight attention on the performance and results of government activities by requiring that all federal agencies measure and report on the results of their activities annually. Agencies are required to develop a strategic plan that sets goals and objectives for at least a 5-year period, an annual performance plan that translates the goals of the strategic plan into annual targets, and an annual performance report that demonstrates whether the targets are met. The Committee on Science, Engineering, and Public Policy (COSEPUP) of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine has addressed the issue of measuring and evaluating research in compliance with the requirements of GPRA.

---

*Reprinted with permission from Evaluating Federal Research Programs: Research and the Government Performance and Results Act, pp. 1–23 (Executive Summary and Statement of the Problem). Copyright 1999 by the National Academy of Sciences. Courtesy of the National Academy Press, Washington, DC.*

COSEPUP recognizes the opportunities and challenges that GPRA presents for agencies that invest in research. GPRA offers those agencies the opportunity to communicate to policy-makers and the public the rationale for and results of their research programs. At the same time, GPRA presents substantial challenges to the agencies.

During the course of this study, COSEPUP held several workshops. In these workshops and in other input to the committee, we have heard two distinct and conflicting viewpoints on approaches to measuring basic research. One is that it should be possible to measure research, including basic research, annually and provide quantitative measures of the useful outcomes of both basic and applied research. The other is that, given the long-range nature of basic research, there is no sensible way to respond to the GPRA annual measurement requirement and that the best that can be done is to provide measures that appear to respond but in fact are essentially meaningless, such as a list of an agency's top 100 discoveries of the preceding year.

COSEPUP's view, spelled out in more detail in what follows, is different from both those viewpoints. In essence, our report takes two strong positions. First, the useful outcomes of basic research cannot be measured directly on an annual basis, because the usefulness of new basic knowledge is inherently too unpredictable; so the usefulness of basic research must be measured by historical reviews based on a much longer timeframe. Second, that does not mean that there are no meaningful measures of performance of basic research while the research is in progress; in fact, the committee believes that there are meaningful measures of quality, relevance, and leadership that are good predictors of eventual usefulness, that these measures can be reported regularly, and that they represent a sound way to ensure that the country is getting a good return on its basic research investment.

The problem of reporting on applied research is much simpler: it consists of systematically applying methods widely used in industry and in some parts of government. For example, an applied research program usually includes a series of milestones that should be achieved by particular times and a description of the intended final outcomes and their significance. Periodic reporting can indicate progress toward those milestones.

The remainder of this executive summary provides a more in-depth description of COSEPUP's conclusions and recommendations regarding how to evaluate federal research programs relative to GPRA. It also addresses coordination among federal research programs and human-resource issues. COSEPUP concludes that both basic research and ap-

plied research programs<sup>1</sup> can be meaningfully evaluated on a regular basis. For the applied research programs of the mission agencies, specific practical outcomes can be documented and progress toward their achievement can be measured annually. For example, if the Department of Energy adopted the goal of producing cheaper solar energy, it could measure the results of research directed toward decreasing the cost of solar cells; this applied research project would be evaluated annually against specific measurable milestones. However, the practical outcomes of basic research in science and engineering can seldom be identified while the research is in progress. Basic research has annual results that can be meaningfully evaluated, but these evaluations often do not give even a hint of ultimate practical outcomes.

History tells us unmistakably that by any measure, the benefit to the United States for leadership in basic research is extremely high—lives saved, inventions fostered, and jobs and wealth created. History also shows us how often basic research in science and engineering leads to outcomes that were unexpected or took many years or even decades to emerge. COSEPUP strongly believes that measures of the practical outcomes of basic research usually must be retrospective and historical and that the unpredictable nature of practical outcomes is an inherent and unalterable feature of basic research. For example, pre-World War II basic research on atomic structure contributed to today's Global Positioning System, an outcome of great practical and economic value, but, attempts to evaluate a year's worth of that early research even if they demonstrated high quality and world leadership, would have contained no hint of this particular outcome.

Since we cannot predict the ultimate practical outcomes of basic research, we must find ways to ensure that the basic research programs that the nation funds generate the kinds of knowledge that have given us great practical benefits in the past. To do that, we must find ways to measure the quality of our current research programs, their contributions to our world leadership in the relevant fields, and their relevance to agency goals and intended users.

World leadership is an important measure. In an earlier report (COSEPUP, 1993), COSEPUP recommended that, for the sake of the nation's well-being, the United States be among the leaders in all major fields of science and pre-eminent in selected fields of national importance. That is because a nation must be performing research at the forefront of a field if it is to understand, appropriate, and capitalize on current advances in that field, no matter where in the world they occur. New

knowledge has value to nations where highly educated people performing cutting-edge research in the field of discovery can make use of the new knowledge when practical outcomes appear possible.

The people best qualified to evaluate basic or applied research are those with the knowledge and experience to understand its quality, and, in the case of applied research, its connection to public and agency goals. Evaluating basic research requires substantial scientific or engineering knowledge. Evaluating applied research requires, in addition, the ability to recognize its potential applicability to practical problems.

With those considerations in mind, COSEPUP has reached six conclusions and offers six recommendations regarding the evaluation of federally supported research programs.

**Conclusion 1:** Both applied research and basic research programs supported by the federal government can be evaluated meaningfully on a regular basis.

**Conclusion 2:** Agencies must evaluate their research programs by using measurements that match the character of the research. Differences in the character of the research will lead to differences in the appropriate timescale for measurement, in what is measurable and what is not, and in the expertise needed by those who contribute to the measurement process.

For applied research programs, progress toward specified practical outcomes can usually be measured annually by using milestones and other fairly standard approaches common in industry and in some parts of the federal government. For basic research, in contrast, progress toward practical outcomes cannot be measured annually, and attempts to measure such progress annually can in fact be harmful. Basic research progress can be reported annually in terms of quality, leadership, and relevance to agency goals, but practical outcomes can be measured only against a far longer historical perspective. In practical terms, because quality, leadership, and relevance will usually change slowly, the GPRA annual-reporting requirement can usually be met by minor updating of full evaluations that are done in a more flexible timeframe. There is a much greater chance that important events will take place in subfields, because of either scientific events or funding changes, so subprogram changes should constitute much of the updating.

Different expertise is required for measuring the worth of applied research and the worth of basic research. Measuring both requires technical and scientific knowledge, but applied research entails some factors

that basic research does not, such as ultimate usability, so the input of potential users is required. That leads to our next conclusion.

**Conclusion 3:** The most effective means of evaluating federally funded research programs is expert review. Expert review—which includes quality review, relevance review, and benchmarking—should be used to assess both basic research and applied research programs.

Expert review is widely applied—used, for example, by congressional committees, by other professions, by industry boards, and throughout the realm of science and engineering—to answer complex questions through consultation with expert advisers. It is useful in helping an agency answer three kinds of questions of particular relevance to GPRA:

What is the quality of the research program—for example, how good is current research work compared with other work being conducted in the field?<sup>2</sup> This question is best answered by reviewers who are sufficiently expert in the field being assessed to perform a quality review. This approach is traditionally called peer review. Peer review is commonly applied to projects, but here we are applying it to programs. The talent, objective judgment, and experience of these experts, or peers, are paramount and should be the criteria for their selection.

Is the research program focused on the subjects most relevant to the agency mission? Another form of expert review is relevance review, in which potential users, joined by experts in related fields, evaluate the relevance of research to agency goals—is the research on subjects in which new understanding could be important in fulfilling the agency's mission? In reviewing the relevance of a program, a panel would assess the appropriateness of the direction of the research to the agency mission and its potential value to intended users.

Is the research being performed at the forefront of scientific and technological knowledge? This is a relevant question for many programs, but it is particularly important for whole fields and subfields being supported. Evaluations of fields and subfields is best done through international benchmarking by a panel of experts who have sufficient stature and perspective to assess the international standing of research.

For agencies whose missions include a specific responsibility for basic research—such as the National Science Foundation in broad fields of science and engineering, the National Institutes of Health in fields related to health, or the Department of Energy in high-energy physics—world leadership in a field can itself be an agency goal. That is equally true for mission agencies, such as Department of Defense (DOD) but in

more focused ways. For example, DOD can take as a goal world leadership in basic materials research relevant to its mission. Once such a goal is established, the usual measures of quality and leadership should be applied.

**Conclusion 4:** The nation cannot benefit from advances in science and technology without a continuing supply of well-educated and well-trained scientists and engineers. Without such a flow, the capability of an agency to fulfill its mission will be compromised. Agencies must pay increased attention to their human-resource requirements in terms of training and educating young scientists and engineers and in terms of providing an adequate supply of scientists and engineers to academe, industry, and federal laboratories.

Federal agencies that support research and exploit its results are able to do so because the education and training programs of the universities, in the course of performing much of that research, and the federal laboratories provide a continuing flow of qualified scientists and engineers. Even though section 1115(a)(3) of GPRA requires agencies to describe the human resources required to meet their performance goals, few agencies describe the importance of human resources or propose ways to ensure their adequacy in their strategic or performance plans.

**Conclusion 5:** Mechanisms for coordinating research programs in multiple agencies whose fields or subject matters overlap are insufficient.

It is common and valuable for agencies to approach similar fields of research from different perspectives. Indeed, this pluralism is a major strength of the U.S. research enterprise. But, better communication among agencies would enhance opportunities for collaboration, help keep important questions from being overlooked, and reduce instances of inefficient duplication of effort. Present mechanisms need strengthening.

**Conclusion 6:** The development of effective methods for evaluating and reporting performance requires the participation of the scientific and engineering community, whose members will necessarily be involved in expert review.

The researchers who work in agency, university, and industrial laboratories are the people who perform and best understand the research programs funded by the federal government. Many researchers contribute substantial time and effort to reviewing papers submitted for publication, grant applications, and program proposals, yet few of them are aware of GPRA, its objectives, and its mandates. Increased contact with

and advice from the broader scientific and engineering community regarding the methods of determining and reporting quality and regarding the leadership position of agency research programs and the relevance of research to agency missions can benefit the GPRA process.

On the basis of those conclusions, COSEPUP offers the following recommendations:

**Recommendation 1:** Because both applied research and basic research can be evaluated meaningfully on a regular basis and are vital to research and mission agencies, research programs should be described in strategic and performance plans and evaluated in performance reports.

The performance of research is critical to the missions of many federal agencies. Therefore, a full description of an agency's goals and results, which is a principal objective of GPRA, must contain an evaluation of research activities and their relevance to the agency's mission.

**Recommendation 2:** For applied research programs, agencies should measure progress toward practical outcomes. For basic research programs, agencies should measure quality, relevance, and leadership. In addition, agencies should conduct periodic reviews of the overall practical outcomes of an agency's overall past support of applied and basic research. The use of measurements needs to recognize what can and cannot be measured. Misuse of measurement can lead to strongly negative results; for example, measuring basic research on the basis of short-term relevance would be extremely destructive to quality work.

Because the evaluation of applied research is directly connected to practical outcomes, whereas the evaluation of basic research is in terms of quality, relevance, and leadership, which ultimately lead to practical outcomes, there might be a tendency to bias an agency's overall research program toward applied research at the expense of basic research. This should be avoided, and a proper balance should be maintained.

**Recommendation 3:** Federal agencies should use expert review to assess the quality of research they support, the relevance of that research to their mission, and the leadership of the research. Expert review must strive for balance between having the most knowledgeable and the most independent individuals serve as members. Each agency should develop clear, explicit guidance with regard to structuring and employing expert review processes.

The most effective way to evaluate research programs is by expert review. The most commonly used form of expert review of quality is peer review. This operates on the premise that the people best qualified to judge the quality of research are experts in the field of research. This premise prevails across the research spectrum, from basic research to applied research. A second form of expert review is relevance review, in which potential users and experts in other fields or disciplines related to an agency's mission or to the potential application of the research evaluate the relevance of research to the agency's mission. A third form of expert review is benchmarking, in which an international panel of experts compares the level of leadership of a research program relative to research being performed worldwide.

**Recommendation 4:** Both research and mission agencies should describe in their strategic and performance plans the goal of developing and maintaining adequate human resources in fields critical to their missions both at the national level and in their agencies. Human resources should become a part of the evaluation of a research program along with the program's quality in terms of research advancement, relevance in terms of application development, and leadership in terms of the ability to take advantage of opportunities when they arise.

In early drafts of strategic and performance plans, agencies have generally omitted discussions of education and training, which are fundamental to the ability of agencies to fulfill their missions. The goal of developing and maintaining adequate human resources in fields critical to their missions should be supported by plans that produce that outcome. The nation cannot benefit from advances in science and technology without a continuing supply of well-educated and well-trained scientists and engineers. In addition, in the absence of such a flow, the capability of an agency to fulfill its mission will be compromised and the knowledge learned and technology developed will be lost.

**Recommendation 5:** Although GPRA is conducted agency-by-agency, a formal process should be established to identify and coordinate areas of research that are supported by multiple agencies. A lead agency should be identified for each field of research and that agency should be responsible for assuring that coordination occurs among the agencies.

It is common and valuable for multiple agencies to approach similar fields of research from different perspectives. Indeed, this pluralism is a major strength of the U.S. research enterprise. However, better communication among agencies would enhance opportunities for collaboration,

help to keep important questions from being overlooked, and reduce instances of inefficient duplication of effort. A single agency should be identified to serve as the focal point for each particular field of research so that all significant supported fields are covered. Information regarding support for that field should be provided to all the agencies involved in it so that they can adjust their efforts to ensure that the field is appropriately covered. Agencies should use benchmarking, which affords the opportunity to look across fields, in their efforts to understand the status of a particular field of research.

**Recommendation 6:** The science and engineering community can and should play an important role in GPRA implementation. As a first step, they should become familiar with agency strategic and performance plans, which are available on the agencies' web sites.

The researchers who work in agency, university, and industrial laboratories are the people who perform and best understand the research programs funded by the federal government. Many researchers contribute substantial time and effort to reviewing papers submitted for publication, grant applications, and program proposals, but few of them are aware of GPRA. Their greater involvement in implementing GPRA would be beneficial to the country. Increased contact with and advice from the broader scientific and engineering community regarding both the quality and the leadership position of agency research programs and the relevance of the research to agency missions can benefit the GPRA process.

COSEPUP intends to address mechanisms and guidelines for implementing these recommendations in workshops and meetings with representatives from federal agencies, Congress, OMB, and oversight bodies. Given the diverse portfolio of research conducted by federal agencies and the urgency of addressing the question of how basic research can be evaluated in the context of GPRA, the level of detail and specificity needed in designing procedures and guidelines for implementation was beyond the scope of this report.

The Government Performance and Results Act provides an opportunity for the research community to ensure the effective use of the nation's research resources in meeting national needs and to articulate to policy-makers and the public the rationale for and results of research. We believe that our recommendations can assist federal agencies in complying with GPRA.

## Statement of the Problem

### *GPRA and Research*

In 1993, Congress passed the Government Performance and Results Act (GPRA) with broad bipartisan support. The law is part of a set of budget-reform measures intended to increase the effectiveness and efficiency of government. Both the General Accounting Office (GAO) and the Office of Management and Budget (OMB) testified in favor of the bill, and the President's National Performance Review advocated its implementation. Unlike several predecessor systems (program planning and budgeting, management by objectives, and zero-based budgeting), GPRA is not an executive branch initiative but rather a congressional mandate. It has received a high level of attention in both the Senate and the House of Representatives.

The specific goal of GPRA is to focus agency and oversight attention on the outcomes of government activities—the results produced for the American public. The approach is to develop measures of outcomes that can be tied to annual budget allocations. To that end, the law requires each agency to produce three documents: a strategic plan, which sets general goals and objectives over a minimal 5-year period; a performance plan, which translates the goals of the strategic plan into annual targets; and a performance report, which demonstrates whether the targets were met. Agencies delivered the first required strategic plans to Congress in September 1997 and the first performance plans in the spring of 1998. Performance reports are due in March 2000. The law calls for strategic plans to be updated every 3 years and the other documents annually.

The general principles of GPRA have been implemented by many state governments and in other countries (for example, Canada, New Zealand, and the U.K.), but implementation by the U.S. federal government is the largest scale application of the concept to date and somewhat different. Over the last 5 years, various states have tried to develop performance measures of their investments. With respect to performance measures of science and technology activities, states tend to rely on an economic-development perspective with measures reflecting job creation and commercialization. Managers struggle to define appropriate measures, and level-of-activity measures dominate their assessments.<sup>3</sup> With respect to other countries, our limited review of their experiences showed that most are struggling with the same issues that the United States is concerned with, notably how to measure the results of basic research.

Not every aspect of the system worked perfectly the first time around in the United States. Some agencies started the learning process earlier and scaled up faster than others. OMB allowed considerable agency experimentation with different approaches to similar activities, waiting to see what ideas emerged. The expectations of and thus the guidance from the various congressional and executive audiences for strategic and performance plans have not always been the same and that has made it difficult for agencies to develop plans agreeable to all parties. Groups outside government that are likely to be interested in agency implementation of GPRA have not been consulted as extensively as envisioned. There is general agreement that all relevant parties should be engaged in a continuing learning process, and there are high expectations for improvement in future iterations.

The development of plans to implement GPRA has been particularly difficult for agencies responsible for research activities supported by the federal government. A report by GAO (GAO, 1997) indicates that measuring performance and results is particularly challenging for regulatory programs, scientific research programs, and programs that deliver services to taxpayers through third parties, such as state and local governments.

### *Findings from Workshops*

From January through June 1998, COSEPUP held a series of workshops to gather information about the implementation of GPRA. The first workshop, cosponsored with the Academy Industry Program, focused on the approaches that industry uses to develop strategic plans and performance assessments. Industry participants emphasized the importance of having a strategic plan that clearly articulates the goals and objectives of the organization. One of the industry participants said that the objective of their industrial research is “knowledge generation with a purpose.” The industry representative indicated that the company must first support world-class research programs that create new ideas; second, relate the new ideas to an important need within the organization or project; and third, build new competence in technologies and people. With respect to performance assessment, many industry participants noted that results of applied research and development programs are more easily quantified than results of basic research. However, even though they might not be able to quantify results of basic research, they nonetheless support it because they believe it important to their business; investments in basic research do pay off over time.<sup>4</sup>

With respect to assessing basic research, industry representatives indicated that they must rely on the judgment of individuals knowledgeable about the content of the research and the objectives of the organization to evaluate the results of such efforts. Some industry participants stressed the importance of giving careful consideration to any metrics one adopts—whether in industrial or government research. It is important to choose measures well and use them efficiently to minimize non-productive efforts. The metrics used also will change the behavior of the people being measured. For example, in basic research, if you measure relatively unimportant indicators, such as the number of publications per researcher instead of the quality of those publications, you will foster activities that may not be very productive or useful to the organization. A successful performance assessment program will both encourage positive behavior and discourage negative behavior. Metrics must be simple, not easily manipulated, and drive the right behavior. Most industry R&D metrics are more applicable to assessing applied research and technology development activities in the mission agencies.

The second COSEPUP workshop focused on the strategic and performance plans of 10 federal agencies: the Department of Defense, the Department of Energy, the Department of Transportation, the Department of Agriculture, the National Aeronautics and Space Administration, the National Institutes of Health, the National Science Foundation, the Environmental Protection Agency, the National Institute of Standards and Technology, and the National Oceanic and Atmospheric Administration. As might be expected, most of these organizations use different approaches to translate the goals in their strategic plans into performance goals for scientific and engineering research. Some agencies use qualitative, others quantitative, and still others, a combination of qualitative and quantitative measures. There was a strong consensus among the agencies that the practical outcomes of basic research cannot be captured by quantitative measures alone. Agency representatives generally agreed that progress in program management and facility operation can be assigned quantitative values.

Agencies with long-term targeted research goals have generally translated them into short-term milestones that can be achieved within a 2-year time horizon for performance planning and reporting. Agencies that seek advances in knowledge in broad fields rather than targeted ones, have not used the milestone approach to performance planning and reporting.

Some agencies have had difficulty in implementing GPRA. When preparing GPRA strategic and performance plans, some agencies are more likely than others to highlight research activities. The major variable is the magnitude of research relative to the agency's other activities. Submersion of research within large agencies makes it impossible for an integrated view of the federal science and technology investment to emerge through the GPRA process and is therefore a matter of concern for COSEPUP.

The performance plans of the agencies tend to emphasize short-term applied research with practical outcomes. Some participants expressed concern that this emphasis would skew funding away from long-term research that is difficult to measure against annual milestones.

Some participants indicated that a desirable result of GPRA would be to increase teamwork among the agencies, as well as to improve communication between research agencies and oversight entities, including Congress, OMB, and GAO. Another theme that recurred throughout the workshop was that the research community has a low level of awareness and is not strongly involved in the GPRA process.

The education and training of graduate and undergraduate students are among the most important duties and durable legacies of the research agencies. Yet human resources was not thoroughly identified or addressed in most agencies' performance plans.

Peer review was identified as the primary method for assessing the quality of research. However, the process by which peer review is applied varies widely among the agencies. Peer review of projects, grants, and contracts differs from peer review of programs and of intramural and extramural research. Those differences led COSEPUP to hold a third workshop focused on peer review and other methods for evaluating research.

In its third workshop, COSEPUP discussed the various methods available for evaluating research. As a result of that workshop and other discussions, COSEPUP found that the following methods are currently available for analyzing research:

- Bibliometric analysis
- Economic rate of return
- Peer review
- Case study

- Retrospective analysis
- Benchmarking

Each of these methods is briefly described below.<sup>5</sup> The pros and cons associated with each technique are summarized in Table 1, later in this chapter.

### Bibliometric Analysis<sup>6</sup>

A technique known as bibliometric analysis, which includes publications, citations, and patent counts, is based on the premise that a researcher's work has value when it is judged by peers to have merit. A manuscript is published in a refereed journal only when expert reviewers and the editor approve its quality; a published work is cited by other researchers as recognition of its authority; and a published work is cited as evidence by a company applying for a patent. By extension, the more times a work is cited, the greater its merit. The primary benefit of bibliometric analysis is its quantitative nature. Furthermore, it correlates well (approximately 60% in one study) with peer review when both methods are used.

The primary argument against bibliometric analysis is that bibliometric measurements treat all citations as equally important. However, many citations refer to routine methods or statistical designs, modifications of techniques, or standard data or even refute the validity of a paper. Other problems are caused by citing the first-named author of a publication when the customs that determine the order in which authors are listed vary by fields. In addition, different mores among research communities—whether particular disciplines or countries—can skew results when they are used comparatively (for example, far fewer outlets are available for Russian publications than for U.S. publications). Furthermore, in emphasizing counts, researchers are apt to take actions that artificially increase the number of citations they receive or reduce their research in fields that offer less opportunity of immediate or frequent publication or in critical related fields (such as education) that do not offer publication opportunities.

### Economic Rate of Return

In recent years, economists have developed a number of techniques to estimate the economic benefits (such as rate of return) of research.

**Table 1: Current Methods Used for Evaluating Research**

<b>Methods</b>	<b>Pro</b>	<b>Con</b>
Bibliometric analysis	Quantitative; useful on aggregate basis to evaluate quality for some programs and fields	At best, measures only quantity; not useful across all programs & fields; comparisons across fields or countries difficult; can be artificially influenced
Economic rate of return	Quantitative; shows economic benefits of research	Measures only financial benefits, not social benefits (such as health-quality improvements); time separating research from economic benefit is often long; not useful across all programs and fields
Peer review	Well-understood method and practices; provides evaluation of quality of research and sometimes other factors; already an existing part of most federal-agency programs in evaluating the quality of research projects	Focuses primarily on research quality; other elements are secondary; evaluation usually of research projects, not programs; great variance across agencies; concerns regarding use of “old boy network;” results depend on involvement of high-quality people in process
Case studies	Provides understanding of effects of institutional, organizational, and technical factors influencing research process, so process can be improved; illustrates all types of benefits of research process	Happenstance cases not comparable across programs; focus on cases that might involve many programs or fields making it difficult to assess federal-program benefit
Retrospective analysis	Useful for identifying linkages between federal programs and innovations over long intervals of research investment	Not useful as a short-term evaluation tool because of long interval between research and practical outcomes
Benchmarking	Provides a tool for comparison across programs and countries	Focused on fields, not federal research programs

The primary benefit of this method is that it provides a metric of research outcomes. However, there are a number of difficulties. In particular, the American Enterprise Institute (AEI, 1994) found that existing economic methods and data are sufficient to measure only a subset of important dimensions of the outcomes and impacts of fundamental science. Economic methods are best suited to assessing mission-agency programs and less-well suited to assessing the work of fundamental research agencies, particularly on an annual basis. Furthermore, economists are not able to estimate the benefit-to-cost ratio “at the margin” for fundamental science (that is, the marginal rate of return—or how much economic benefit is received for an additional dollar investment in research), and it is this information that is needed to make policy decisions. Finally, the time that separates the research from its ultimate beneficial outcome is often very long—50-some years is not unusual.

### Peer Review<sup>7</sup>

Peer review is the method by which science exercises continuous self-evaluation and correction. It is the centerpiece of many federal agencies' approach to evaluating proposed, current, and past research in science and engineering.

Peer review, like all human judgments, can be affected by self-interest, especially the favoritism of friendship and the prejudice of antagonism. However, those distortions can be minimized by the rigor of peer selection, the integrity and independence of individual reviewers, and the use of bibliometric analysis and other quantitative techniques to complement the subjective nature of peer review.

Peer review is not equally appropriate across the wide span of research performed by federal agencies. We might visualize at one end of the spectrum the fundamental, long-term projects whose ultimate outcomes are unpredictable and at the other end programs of incremental or developmental work whose results are easier to predict within fairly narrow time limits. Projects of the latter type can often be evaluated in a rigorously quantifiable fashion by appropriate metrics. It is for the former kind of research, whose results are not easily quantified, especially while the work is in progress, that peer review of quality and leadership is required and generally effective. Agency managers have the responsibility of designing review techniques that suit the nature of each individual research program being evaluated.

## Case Studies

Historical accounts of the social and intellectual developments that led to key events in science or applications of science illuminate the discovery process in greater depth than other methods. The chief advantage of case studies is that they can be used to understand the effects of institutional, organizational, and technical factors on the research process and can identify important outcomes of the research process that are not purely intellectual, such as the collaboration of other researchers, the training of young researchers, and the development of productive research centers. Difficulties of case studies are that they can be expensive, and that the validity of the results and conclusions depends on the objectivity, investigative skills, and scientific knowledge of the persons doing them.

## Retrospective Analysis

Retrospective analyses are related to case studies in that they also try to reconstruct history; however, they focus on multiple scientific or technological innovations rather than just one. The goal is to identify linkages between innovations and particular types of antecedent events (usually either funding or research). Such analysis is usually done by a panel of experts or investigators. This method is most appropriate for assessing a particular type of accountability question (for example, impact of National Science Foundation funding on mathematics research). The primary disadvantage of this type of analysis is that it takes a long time to conduct and thus is not useful as a tool to provide short-term evaluations for improving research policy and management.

## Benchmarking<sup>8</sup>

As noted earlier, maintaining leadership across the frontiers of science is a critical element of the nation's investment strategy for research (COSEPUP, 1993). The question addressed here is, whether an agency's or the nation's research and educational programs are at the cutting edge? This assessment is made by a panel of international and national academic and industrial experts in a given field and in related fields on the basis of available quantitative and qualitative data. COSEPUP has conducted a number of experimental efforts on benchmarking the United

States' position in selected fields. Programs can be benchmarked in a similar fashion.

## Endnotes

1. For purposes of this study, program refers to a set of activities focused on a particular area that can include multiple projects with different risks, time horizons, and outcomes.
2. There are at least two aspects of quality—one absolute and one relative. The absolute aspects are related to the quality of the research plan, the methods by which it is being pursued, its role in education when conducted at a university, and the importance of its results to its sponsor, either obtained or expected. The relative aspects pertain to its leadership at the edge of an advancing field. Although the leadership aspect is generally important, the results might in some cases be of great importance to an agency albeit not at the leading edge of a field.
3. For more information regarding individual states see <http://www.gsu.edu/~padjem/projects.html>. [G-14]
4. For additional information on corporate experience in assessing research and its applicability to federal research, see Commission on Physical Sciences, Mathematics, and Applications, (1995) *Research Restructuring and Assessment*, National Academy Press, Washington, D.C.
5. These descriptions were adapted from the National Science and Technology Council's (NSTC) *Assessing Fundamental Science*, 1996.
6. Small, Henry G. "A Co-Citation Model of a Scientific Specialty: A Longitudinal Study of Collagen Research" *Social Studies of Science*, Vol. 7 (1977), 139-66. Anderson, Richard C., F. Narin, Paul McAllister "Publication Ratings versus Peer Ratings of Universities" *Journal of the American Society for Information Science March* (1978) 91-103.
7. For additional information on peer review, see Atkinson, Richard C. and William A. Blanpied, Peer Review and the Public Interest, *Issues in Science and Technology*, vol 1. no. 4, 1985; Bozeman, B. and J. Melkers, "Peer Review and Evaluation of R&D Impacts," *Evaluating R&D Impacts*, Kluwer Academic Publishers, Norwell, Mass., (1993) 79-98; Cole, J. and S. Cole, *Peer Review in the National Science Foundation*, Washington, D.C.: National Academy Press, 1981; GAO, *Peer Review; Reforms Needed to Ensure Fairness in Federal Agency Grant Selection*, June 1984.
8. See COSEPUP, 1997 and COSEPUP, 1998.