SAND95-2425C
CONF-9511140--Summ.
CONF-9509202--Summ.

R&D Evaluation Workshop Report

U.S. Department of Energy, Office of Energy Research

September 7-8, 1995

Prepared for the Office of Basic Energy Sciences
by
Sandia National Laboratories
Energy Policy and Planning Development

October 30, 1995

Gretchen Jordan, Ph.D.
(703) 247-3611

DRAFT

Distribution:
Martha Krebs
Iran Thomas
Al MacLachlan
Bill Valdez
Attendance List
Invitation List

This work was supported by the United States Department of Energy under Contract DE-AC04-94AL85000.

Comments to Gretchen Jordan, 703-247-3611, gbjorda@Sandia.gov, fax 703-516-4418 by Thursday, November 16.

TABLE OF CONTENTS

NOTE: Appendices not mailed to those attending. Call to request, if needed.

R&D Evaluation Workshop Report
September 7- 8, 1995
U.S. Department of Energy, Office of Energy Research

Executive Summary

- The workshop, sponsored by the U.S. Department of Energy (DOE) and Office of Energy Research (ER), met its primary objective - to promote discussion among evaluation experts and DOE research managers on developing approaches for assessing the impact of DOE basic research upon the DOE energy mission, applied research, technology transfer, the economy and society.

- R&D evaluation is a difficult yet timely issue. This was evidenced by the press the workshop received. "DOE Researchers See a Challenge in Developing Project Assessments" appeared in Inside Energy (September 11, 1995), and "Evaluating Basic R&D is Hairy But Necessary, DOE Forum Finds" appeared in the Federal Technology Report (September 14, 1995).

- The two day session was attended by 80 persons, including both evaluation experts and a cross section of the research community. Participants found that it was a valuable workshop which brought people together to find common ground. Many gained a better grasp of the complexity of this issue, new ideas came out of the workshop, and the participation of DOE management demonstrated commitment to this issue. The speed at which ER will implement measurement of appropriate performance measures for its programs was increased by the workshop.

- In a questionnaire, most participants concluded that the information learned at the workshop will be valuable in their own work environment. On a scale of 1-5, with 5 being excellent, overall performance was rated 4. Respondents were, in general, satisfied with the performance of the facilitators and experts in attendance, as well as the introductory panel. Very good marks were given for the Breakout Sessions on Thursday, however, respondents were more satisfied with their performance on Friday.

- During the workshop participants learned about what is already being done to address the issue of R&D evaluation. Discussion with evaluation experts on innovative evaluation methods was begun, particularly in the area of case studies. Most of the discussion, however centered on the audience and purpose of evaluation and what measures were appropriate.

- Although participants recognize the need for more accountability and the need to communicate the value and impact of ER programs, there is concern about increased bureaucracy and the burden of data collection. There is concern that information be collected and presented with care because it can be misinterpreted and misused.

- What ER is trying to do is very complex. ER must measure for multiple purposes (communication to public and program improvement) and multiple expected audiences, in varied context, multiple time frames, and for several types of R&D programs.

- ER should account for the value of innovation and risk. It must be recognized that a climate which encourages discovery and innovation can only be achieved by a bold willingness to underwrite risk. It is also important to distinguish between the value of research that is badly designed (which has little or no value) and research that is well-designed (which always has value, whether it yields "positive" or "negative" results).

- It is important to determine what level or unit to evaluate and report - office, program or project. The preferred unit was unclear during the workshop and no agreement was reached. The unit and the measurement will be different for different evaluation purposes and target audiences.

- Determining the evaluation methods will be relatively straight forward once ER has decided on the minimum set of measures needed (keeping in mind the potential data collection burden). A balance of quantitative and qualitative data is necessary for the chosen "cafeteria" of measures. There is no universal solution and each method has strengths and weaknesses.

- The next step is to have more clear direction from ER management on who the audience is and what measures are likely to satisfy those audiences. In a facilitated session managers could develop a "performance framework". In determining the performance framework, ER should start with the DOE strategic plan and the existing set of ER measures as defined in Contract reform documents. The framework should include the universities and facilities in the discussion and be the basis for an integrated measurement system.

- The recommendation was made to hold a workshop in January 1996 with the September participants to continue the discussion on evaluation of the impact of ER programs. The first morning, presentations would develop a common understanding of performance measurement planning and selected evaluation methods. Then breakout sessions would confirm and further define the measures provided by ER management, review current data collection for what can be discarded and what must be added, and discuss evaluation strategy and methods for collecting the additional data required.

R&D Evaluation Workshop Report
September 7- 8, 1995
U.S. Department of Energy, Office of Energy Research

Workshop Objectives, Attendance, and Format

- The objective of the workshop was to promote discussions between experts and research managers on developing approaches for assessing the impact of DOE's basic energy research upon the energy mission, applied research, technology transfer, the economy, and society. The purpose of this impact assessment is to demonstrate results and improve ER research programs in this era when basic research is expected to meet changing national economic and social goals.

- Attending the workshop were the ER Director, the Deputy UnderSecretary for R&D Management, ER Associate and Division Directors, representatives of DOE laboratories, field offices, and applied research and technology partnership and policy offices, and members of the Panel on the Value of Basic Research. The workshop was moderated by the Director (Acting) of the ER Office of Basic Energy Sciences (OBES). Project managers and the Technology Steering Group of the OBES Center of Excellence for the Synthesis and Processing of Advanced Materials were present. Special guest at the Thursday evening reception was Mary Good, UnderSecretary of the U.S. Department of Commerce.

- R&D evaluation experts present included: Francis Narin, Susan Cozzens, Harvey Averch, Albert Link, George Teather, Barry Bozeman, J. David Roessner, Ronald Kostoff, and Maria Papadakis. Representatives of other federal agencies and the National Academy of Sciences also attended. A press round table was held during the workshop.

- The workshop approach was to combine R&D evaluation and R&D management expertise to come up with practical suggestions for meeting current needs and for a longer term strategy.

- The workshop was divided into a number of sessions. The first session gave participants a common frame of reference. Panels presented an overview of the current evaluation efforts and an OBES-sponsored Oak Ridge National Laboratory (ORNL) research project that exemplifies the research environment and impacts. Breakout sessions followed with participants placed in groups with mixed evaluation and R&D expertise. Each addressed the same three questions with emphasis on two evaluation methods. There were also breakout sessions the second day, and integrated sessions that gave the participants the chance to share their findings with the entire workshop. The second afternoon a smaller group summarized shared concerns and recommendations.

- The questions addressed were:
  - By what criteria and metrics does Energy Research measure performance and evaluate its impact on the DOE mission and society while maintaining an environment that fosters basic research?

  - What combination of evaluation methods best applies to assessing the performance and impact of OBES basic research? The focus will be upon the following methods: Case studies, User surveys, Citation analysis, TRACES approach, Return on DOE Investment (ROI)/ Econometrics, and Expert panels.

  - What combination of methods and specific rules of thumb can be applied to capture impacts along the spectrum from basic research to products and societal impacts?

R&D Management Presentations

Martha Krebs, Director, Office of Energy Research
"Performance Measurement, Getting it Right"

- ER has several initiatives underway to meet increasing requirements for accountability. It is important to move quickly to meet the challenge of measuring the performance of ER programs in a consistent and effective way. The purpose of measurement is to demonstrate the scientific quality and value of ER research and collaboration and to find predictors of success so ER can make better informed research investments.

Al MacLachlan, Deputy UnderSecretary for R&D Management
"Evaluation of Research Can Be More Than An Archeological Expedition"

- His experience from years in industry is that evaluation of research must be done in context, using predominantly qualitative, not numerical data. The context for research includes the management structure, the infrastructure, the mission, motivation, and opportunities for collaboration.

Iran Thomas, Director (Acting), ER Office of Basic Energy Sciences
Workshop Moderator

- OBES only funds 1 in 15 research proposals, thus program managers must make decisions among them. This competitive selection, along with peer and program review ensures the quality of the science. However, the hope is that this workshop will help improve this decision process and provide insights on how to predict the impact of research without

damaging the environment for quality research. OBES has undertaken several research activities in R&D evaluation to learn better how to measure and communicate impact on the economy and other national goals. The emphasis is on simplicity, less is better. One does not have to look at every leaf to know if the tree is healthy or sick.

Panel Presentations

Example Case - Nickel and Iron Aluminides
Linda Horton, Peter Angelini, Ron A. Bradley, Rod Judkins, and C.T. Liu of Oak Ridge National Laboratory (ORNL)

- Linda Horton, ORNL project manager, and some of her associates presented the history of this OBES program that started as basic research many years ago and has subsequently blossomed with practical applications. In 1980, researchers at ORNL became interested in how grain boundaries in alloys influence strength, and OBES agreed to partly support their work, with the rest of the support coming from laboratory-directed funds. Funding for the research was not tied in any way to expectations for a practical product. The team doing the research was entrepreneurial, however, and unusually aware of developments going on elsewhere. A Japanese language publication, noting that adding small amounts of boron could make these alloys more ductile, provided a crucial link that allowed the researchers to connect very rapidly work at the level of fundamental understanding to the world of economic relevance.

- The results are impressive. The number of papers on intermetallics of all sorts continues to increase from a few per year before the work started to several hundred per year. Thirty invention disclosures and 16 patents have emerged at ORNL, and 12 licenses have been negotiated with industry. Ten cooperative research and development agreements (CRADAs) between ORNL and industry have their origin in intermetallic research or its spinoffs. Today, about fifteen years after the research started, we are seeing the first significant routine industrial applications of the new material. At least twenty prestigious awards and fellowships to ORNL and the researchers as individuals actually preceded the wide-spread use of the nickel and iron aluminide materials.

- The sustained funding of OBES, even after success in the applied research arena, was noted as important to continued success. There were multiple funding sources and success didn't happen quickly. Success occurred at basic and applied levels nearly simultaneously in this case. The key elements of success were a good idea, partnerships in R&D, and a multi disciplinary approach.

Existing Evaluation Efforts in ER: Panel Discussion

ER Office of Basic Energy Sciences

- Iran Thomas noted that in addition to the other activities the workshop will hear about, OBES has a measurement pilot with OMB on facility use. For this ER is tracking three simple measures that will alert us if a problem exists that we need to investigate further. The three measures and the major question each addresses are as follows: Number of hours available compared to total possible hours (are we providing enough funds?), Number of hours scheduled compared to the number available (is there demand for the facility?), and Reliability of the equipment. The later varies across facilities so this will be determined by questionnaire. Common definitions for collection of the first two have been agreed upon. ER is also tracking conventional data such as the number of users and institutions, and is designing a short survey to be completed at the end of use of the facility by each client.

Panel on the Value of Basic Research

- Panel member John Stringer of Electric Power Research Institute spoke for Panel Chairman John Moore (George Mason University). The Panel was organized by the Basic Energy Science Advisory Committee at the request of the Office of Energy Research Director Martha Krebs. The Panel is charged to provide estimates of the impact or value of the research to society in a report due in Spring 1996. Quantitative measures are heavily supplemented by various qualitative assessments. Panel research inquiries include studies of patents that rely on OBES research, a survey to estimate value to industry, and contributions to the education of trained researchers. A series of site visits have been conducted. Of particular concern is looking at what is meant by "value" of basic research and how to assess the value of both success and the null result.

ER Administration Activities

- Phil Stone of ER Science and Technology Affairs, noted that the requirements for performance measurement are coming from many directions. An early effort within the context of DOE "Contract Reform" injects performance measures into contracts with national laboratories as those contracts come up for renegotiation. With input from industry and stakeholders, ER determined that four areas of qualitative measures should become standard across ER. The four areas are: Quality of the basic science, Relevance to DOE missions and national needs, Construction and operation of research facilities that meet user needs and requirements, and Effectiveness and efficiency of research program management.

- The office is also developing a systematic way to gather and aggregate data and to do DOE laboratory appraisals using this basic set of measures. In response to a Galvin Commission Report recommendation the office is also investigating ways to bring uniformity to the lab technical review process.

Innovative Case Studies for OBES

- For OBES, Barry Bozeman and J. David Roessner of the Georgia Institute of Technology have completed three test case studies using the R&D Value Mapping technique. This technique combines case studies and quantitative techniques. Using a common protocol for all cases it is possible to make statistical inferences across cases about what managerial arrangements yield a higher proportion of desired outcomes. They have proposed to OBES that they complete 50-60 case studies using this method, 10 of these randomly selected and the rest with known impact on industry.

- A problem is how to set boundaries on the cases. Synergies exist and are good, but make it difficult to allocate credit to particular researchers or sources of research support. It is often difficult and perhaps futile to be portioning results between DOE and the company, yet it is required. Sometimes, undercounting occurs in the attempt to separate out benefits. Finally, it is hard to find either a cold trail or one with "a thousand flowers".

DOE Office of R&D Management

- Bill Valdez stated that the Office of R&D Management (formerly the Office of Technology Partnerships) is trying to develop a framework to systematically show the value of overall DOE research to the taxpayer. Reports of a few high impact projects and other individual case studies are necessary but not sufficient. The office has three initiatives: Development of an integrated framework for metrics on partnerships that includes the Integrated Technology Transfer System (ITTS) database to track and trend performance indicators like CRADAs; development of customer satisfaction surveys used in an integrated manner to ask the value of products and services such as CRADAs; and a task force to look at M&O contracts and tie them to categories of basic metrics. A desirable fourth effort would be to better communicate the value of cooperative R&D to funders, press and public.

Group Discussion

- Susan Cozzens, principal author of recent publications on the Government Performance and Results Act (GPRA) as it relates to research, stated that the GPRA does not just ask for measurement of program outcomes. It requires definition of strategic goals, with measures

used to demonstrate annual progress toward those goals. It requires reporting at a broad level, that is at the ER level, not at the level of the ORNL example case.

- Group discussion pointed out differences in approach to evaluation of research. Scientists want to be more accurate and have detailed studies like the OBES case studies. In contrast, others are satisfied with questionnaires and estimates of economic impacts with no validation. It is not clear what the Congress and its public constituency wants.

- The discovery and innovation processes originate with the seminal insights and creative ideas of our scientific community. The return on our investment in basic science is critically dependent on our selectivity of those explorations, concepts and individuals that we chose to support.

- Longevity of the researcher's experience in the area appears to be valuable so he/she can take advantage of opportunities for success. However, there have been sensational results from young investigators and new research programs. In contrast, programs that are dying tend to be accompanied by a loss of technical leadership, loss of synergism and productivity.

- Some attendees believe that old-fashioned gut-level management is most reliable and stated that one can tell the successful projects by walking down the halls of the research lab.

- It is important that both basic and applied researchers be cognizant of both basic and applied complementary research that relates to their own research. This understanding can often be derived from participation in professional society meetings and symposia, especially those which are neither exclusively basic or applied in program composition. It is also important to encourage the publication of original research findings in peer reviewed scientific journals for several reasons: quality assurance, added value to the published findings, results broadly disseminated and subjected to the maximum amount of scrutiny by the entire scientific community.

- The direction and creativity of both basic and applied research are dynamic evolutionary processes. They both require freedom to make decisions at the project level. Successful research will always be evolutionary in nature, with approaches and solutions being continuously created and either eliminated or set aside.

Breakout and Integrating Group Discussions

Focus of Breakout Session Discussion

- The groups began discussion with the three questions presented on page 2 above. Thursday's integrating session narrowed the focus of discussion for the second day breakout sessions to what measures and methods are needed for "Communication - by ER and OBES -to various stakeholders (DOE, OSTP, OMB, Congress) - of
  (1) impact of basic research on more applied research
  (2) impact of basic research on industry/economy and society
  (3) effective management of research
  (4) impact/importance of "no" answers and taking risk

Case Study and TRACES Approach Group

- The TRACES approach used in-depth case studies, thus the group considered the two methods as one. Case studies can add credibility to anecdotal stories, can be short and descriptive or in-depth (30 pages) and analytical. Both types can help ER communicate value and improve programs. Case studies can answer, or at least provide some guidance on, subjects such as what worked and why, what is the actual or potential impact, what were costs, what was the influence on "applied" programs, etc. Case studies can be current or retrospective. Managers often gets credit for good management practice for doing case study evaluation even if the resulting insights are not used. Weaknesses of case studies are that they are expensive (often $10,000 or more), time consuming, and bother researchers excessively if poorly managed. Retrospective studies provide more certainty and accuracy but usually provide little data useful for current decisions at the project level.

- In-depth case studies often use multiple lines of evidence, including citation analysis and surveys. Multiple studies, chosen in an organized process using multiple methods and common protocol, could make inferences across studies and investigate predictors of success. In this way, case studies could also help determine what metrics we should be monitoring instead of what we are already gathering.

- Case studies could be done as part of a long-term plan to communicate value and effectiveness. OBES could collect data needed for possible future case studies on a routine basis, and "beef up" existing success stories with in-depth studies. Topics for case studies might be chosen in two ways. ER could look broadly at measures and stories for "markers", or indications of success or problems, and pick a few to investigate further. Or ER could select a "top 10", perhaps major innovations, in a portfolio to demonstrate ER/DOE value.

Surveys and Return on Investment/Econometrics Group

- First ER and OBES must articulate the mission in terms of a set of measurable goals and then determine metrics that help provide evidence of progress toward those goals. Measures are highly user dependent, thus it is essential to define the audience and then speak to that audience be it research managers or policy people. We can't just ask scientists what measures are important. Metrics should be designed to aid in communicating program merits as well as to improve performance.

- A hierarchy of methods may be the best approach. For example, counting CRADAs and licences demonstrates improved transfer of knowledge. Citations and efficiency measures demonstrate significant advance in understanding fundamental science, while cost/benefit analysis links R&D costs and advances with improved national well being. Joint and interdisciplinary interactions demonstrate relevance of research to the mission while peer review demonstrates quality of science. Customer surveys demonstrate the advance of knowledge/support of core competency and may also be able to provide information to help measure the social and economic returns.

- We need to phase evaluation studies and measures to the project life cycle. The time factor will affect what measures one chooses. It is also important to anticipate what would have happened without the research and what effects other R&D had on the impact being investigated, that is, to determine (to the extent possible) a baseline or reference case. There is no one rule book, and in most cases expert judgement is needed.

Citation Analysis and Expert Panels Group

- The focus of evaluation is two fold, to answer the questions, "Is it Good Science?" and "Does it have application to science, the economy, health, the environment and other national priorities?" We must define the programmatic level and stage of a project which we are measuring, and the purpose of the indicators. The "Foresight" process and watching international trends might help identify areas of "relevant" research. Who would do the foresight analysis and how it should fit into the evaluation are topics that need consideration.

- Measuring and communicating the value of research depends partially on addressing communication issues between basic and applied research and the management of research. There needs to be co-operative definition of the contribution of basic and applied research to a solution. ER needs incentives to measure in an integrated manner.

- We can do quantitative retrospective studies such as TRACES and patent citations, with their inherent time lags. We can use measures to communicate how and why general research

areas are chosen. We need a combination of methods, with expert opinion supported by quantitative measures. We should account for the value of innovation and risk, and avoid using the term "failure" for research that results in a "no" answer or "dead end". It must be recognized that a climate which encourages discovery and innovation can only be achieved by a bold willingness to underwrite risk. It is also important to distinguish between the value of research that is badly designed (which has little or no value) and research that is well-designed (which always has value, whether it yields "positive" or "negative" results).

Conclusions or Shared Points of View

- The workshop did a good job of explaining the "why we must measure". Now must have more clear direction from DOE management on who the audience is for the measures, what measures are likely to satisfy those audiences, and the level at which the data is to be collected and reported.

- It is important to determine what level or unit to evaluate and report - office, program or project. The preferred unit was unclear during the workshop and no agreement was reached. The unit and the measurement will be different for different evaluation purposes.

- Information must be collected and presented with care because it can be misinterpreted and misused. We can minimize abuse by preparing for misinterpretation.

- Participants recognized the need for more accountability, but shared concern that there not be increased bureaucracy. Management should try to offset the data collection burden, concentrating efforts on a minimum set and integrating data requests.

- It is particularly difficult to measure and motivate measurement in these times of prolonged ambiguity, with no grand mission shared among DOE offices, or between small and big science.

- DOE is doing a better job than it is communicating. Small changes in way data is collected may allow for better communication of collaboration, joint costs, and impacts.

- What DOE is trying to do is very complex. We must measure for multiple purposes (communication to public and program improvement) and multiple expected audiences, in varied context, multiple time frames, and for several types of R&D programs.

- ER efforts are aimed at the long run goal of having an information performance measurement/management system. Start with an experiment to test various methods, and learn how to measure more efficiently and better.

- In the short term ER needs better communication of the value and impact of research. The long term need is robust data to improve performance. Can the same measures and measurement system provide answers for both purposes?

- Determining the evaluation methods will be relatively straight forward once we have decided on the minimum set of measures needed. A balance of quantitative and qualitative data is necessary for the chosen cafeteria of measures. There is no universal solution and each method has strengths and weaknesses

- Two aspects of research that should be measured and/or communicated better: the positive contributions of "no" answer and the essential ingredient of success - people.

Recommendations for Next Steps

- The recommendation was made to hold a workshop in January 1996 with the September participants to continue the discussion on evaluation of the impact of ER programs. The first morning, presentations would develop a common understanding of performance measurement planning and selected evaluation methods. Then breakout sessions would confirm and further define the measures provided by ER management, review current data collection for what can be discarded and what must be added, and discuss evaluations strategy and methods for collecting the additional data required.

- Before the next workshop, senior managers should meet to answer the basic questions of who is the audience, what is the purpose of performance measurement, and what are priorities for measurement (what is the minimum set of measures). A facilitated session with senior managers could develop more clear answers to these basic questions, resulting in a "performance framework" for ER.

- A good measurement system would have as guiding principles the use of multiple lines of evidence or indicators, so long as these indicators are converging to demonstrate progress toward key goals. Measures must be communicable and measurable, reliable and valid, and are as likely to be qualitative as quantitative. Qualitative measures include issues addressed by expert panels and surveys. There should be a cafeteria of measures from which program mangers choose those that apply to their programs. The strengths and weaknesses and cost/benefit of the measures should be investigated and methods for deleting, adding and modifying measures on a regular basis be set up.

- Measures will vary by time frame (short, mid and long term), by purpose and expected audience (program improvement or reporting to Congress), and by context. Since the audience will change, it would be best to build a robust system that covers most bases.

- ER should try to keep the data collection burden low by determining a small set of key measures, and consider ways to offset the cost of data collection.

- It is important for ER to keep communication open about performance measurement activities within ER (programs and administration), and between ER, the technology offices, and R&D management.

From
Math
Gretchen Jordan-

R&D Evaluation Workshop, DOE Office of Energy Research, September 7, 1995
Holiday Inn Capitol
500 C St. S. W. Washington, D.C.

Evaluation of Research Can Be More Than An Archeological Expedition

by A. MacLachaln

My background is industry, During my years in DuPont, most of which was in research, although I did have several stints in marketing and planning, I saw what I believed was good and bad research. I now think I was wrong in the conclusion that there was good and bad research, rather what I think was the matter was good and bad context for research. This conclusions leads me to the view that evaluation of research per se while potentially quite valuable is not the complete story. We must also look at the context. This is just as true for basic research as all other kinds of research. I for one am extremely skeptical that one can evaluate research in any direct numerical manner. Once more I don't even think you have to.

The title I used for these remarks was to create an image of what I see when someone says they are going to evaluate research. Why? What do you do when you evaluate something? You look at it in its current state or after the fact. That's archeology to me.

Now I'm not knocking archeology. In fact thinking of evaluation of research as archeology can definitely lead to some great incites. I'm going to give a few case studies I was involved in myself to illustrate the point that looking at the context in which research is done should be a fruitful endeavor and may be a better predictor of impact than many other methods. Looking at and evaluating context is also something that is here and now rather than after the fact.

Before we start to evaluate research we have to decide what kind of research we are talking about. Some think it's easy to evaluate incremental research; The kind one does in industry when one makes improvements on current products and processes. On the surface it does seem easy, because you generally get a rapid market response. You can even fool yourself into thinking you can put an accurate dollar value on this kind of research. You can get interesting sounding numbers to show the CEO, but how do you know you got all that was on the table? How sustainable is the advantage you just created? How quickly will your competition follow? How much innovation was really involved? Was it just catch-up, or are you keeping ahead? How much did marketing and manufacturing contribute to the success?

I've been in several businesses where we beat the pants off competition with somewhat inferior products. But, a combination of novelty in our product offerings and spectacular marketing and technical service gave us a growing market share and good profitability for a long period. Does that mean the research was bad? Not at all, it was just a case of reality and using and optimizing all the assets of a business to win in the market place.

I've also been in a business where a customer bought a product from us, improved it with some novel technology and threatened to be more profitable in the market place than we were with this product line. Does that show poor research? Not necessarily, in fact, in this case this was an example of oppressive management that forbid research in the areas that our customer/competitor explored. Our management did this because they thought their research organization was spending too much money and had to be reigned in. This was undoubtedly true, but the approach was wrong and wrecked the business for a long time.

In the latter case I just described, the research was not the problem. The problem was the context in which the research was done.

I can give some very positive examples as well. One of the most interesting times in recent years in

DuPont research was the conversion from the possible ozone destroying refrigerants to the new ozone friendly two carbon materials.

For years we had been one of the major suppliers of these marvelous chemicals. They were perfect for the job of a heat transfer agent in refrigerators. But, scientists had showed there was a growing body of evidence that they were damaging the ozone in the stratosphere. This was just an average business in DuPont. When the need to phase out became clear, we seriously considered abandoning the market. After all, to stay in the business meant massive new investment, an unknowable amount of research, and a mad scramble by very capable competition to win in the new market. The new market was also quite indefinite, because we were already seeing customers eliminating use of the current materials in as many applications as possible. For example, certain printed circuit cleaning operations were being converted to non-fluorocarbon processes. However, it was clear there was still a big refrigerant market, and someone would fill it eventually.

We decided we would try, even though there was recognition the competition would be intense. For the first time in many decades we were in a horse race. The Chairman said very clearly he wanted to win this one and that R&D could have anything within in reason.

We did not hear the reason part. A few ground rules were laid down. They were: We should try to utilize as much of the current plant investment as possible, we should create processes that produced essentially no waste, and we should do this in a way that we can announce withdrawal from the current products as far ahead of the mandated deadlines as we could.

He was true to his word. There was a tremendous sense of excitement in the entire companies research community, Right from the beginning the business leadership acted as though the entire company was available as resources. Pilot plants scheduled for other things in other business lines were requested, and amazingly they were made available. Researchers from all around the company were commandeered, and joined the task force with delight. Corporate research was asked to participate as a full partner, something that had not happened since the days of Nylon scale-up. Within months a whole bunch of novel routes were being investigated simultaneously, theoretical chemists were inventing ways to model plants and thermodynamic behavior of processes and compounds we did not yet have, and universities and national labs were brought in to help us in these and other areas. In four years we went from start to manufacture of the first compounds. Research climbed from a few million dollars a year to more than $30 million at the peak, and we had plants built before we

had fully tested the processes in the pilot stage. Lots of things went wrong. But nothing delayed the schedule. People worked together across all kinds of disciplines. Basic research was integrally coupled to all the information that was needed. DuPont announced its intention to withdraw from the old compounds a year ahead of the mandated schedule. During that whole development time there was an intensive involvement with customers to help them get ready for the very different new compounds. The new routes were able to use large parts of existing facilities, but the new investment still exceeded $500 million, and will eventually top a billion dollars. The processes developed operate at 98 to 99 % yield. We even invented a novel process to quantitatively convert this 1-2% waste back to starting material. A whole new realm of catalyst structures and technology was opened up, which has other potential uses. The total cost of the research was far less than had been originally estimated.

One might well evaluate this whole effort as excellent. It was. But, it was not the research that really carried the day, it was the context in which the research was done. These researchers had been around the company. Some had very good reputations, others not so good. When one looks at the innovations and accomplishments , however one sees no differences among them in this program. Clearly it was the way

things were done. The people were always good, some had just not been plugged in properly until this project came along.

We can't have such a grand challenge every time, but there were many other aspects of this program that were notable contributors to the positive context of the research. It was clear management was really behind the program. There was no game playing or politics. All parts of the company were asked to contribute assets if they had any that would ensure success. Information was always available to anyone. The people in corporate research knew as much about the business goals as they did about the technical status. They were invited, and in fact expected to attend all planning and strategy meetings. No one had to order them to do this. They wanted to because they felt they were welcome and vital partners in the endeavor. What's more, they would decide what work they would do. They were not given just the contingency work. They were invited to do whatever work they thought they were uniquely able to do. Once they committed they were depended on by the business organization to deliver. And deliver they did. They led the computer simulation work for thermodynamic properties of the new materials and their intermediates. This was vital to selection of the compounds for manufacture, and the manufacturing process designs themselves. New catalysts were

invented at an incredible rate. All kinds of problem solving sessions were held with everyone baring their soles. No one cared which organization they came from. Only the science and engineering mattered.

Now , I imagine most of you are thinking, what has this to do with evaluating the effectiveness of basic research? Well, I think it has every thing to do with the subject this meeting. In recent times almost all companies have questioned the value of their basic research organizations, or corporate labs as they are called. They want to understand what they were getting for this investment. And for sure they wanted to get more for that investment.

Let me draw on my experience at DuPont and from my study of several other companies who distinguished themselves over the years through research accomplishments both in direct support of their ongoing businesses and in basic research.

Great companies like General Electric, IBM, DuPont, 3M, Shell, Philips and many others have outstanding records in research, and were known for their great central labs. In recent times, however, almost all companies have questioned the relevancy of their corporate labs. Some have gone so far as to eliminate them. Some have tried to understand the problem and have made adjustments. Those that have chosen this

route will reap great rewards, because they recognize that the ability to hire the best people, and give them the ability to do leading edge science holds the key to long term success. However, these same companies were not happy with the contribution of their corporate labs and recognized they needed to change.

But what was the problem? In DuPont's case I am of course very familiar with the situation. On the surface, things did not look too bad. The publication quality in peer reviewed journals were at the upper end of many of the best universities. The corporate labs had no trouble hiring excellent people. We had the best university consultants, and didn't even have to pay top dollar for them. They often got as many ideas from DuPont scientists as we got from them. The university world complimented company management profusely on their labs. Every time there was a murmur within the company that perhaps the corporate labs should contribute more, our university friends started writing letters to the chairman about the great damage that was about to be done. The management of corporate labs was skillful at stonewalling any ideas for change, especially change that pulled the corporate labs closer to the businesses. When I was in the corporate labs of DuPont in the early 60's there was an unwritten rule that one did not suggest taking a sabbatical to any

business unit in the company. To go to a university was of course highly acceptable.

The management of companies during the 60's and 70's became ever more concerned about the corporate labs. Their costs were escalating rapidly as was all R&D, but the problem was more related to the growing complaints from the businesses. The chief complaint being lack of interest in helping the businesses. This was exacerbated by a growing sense of arrogance and animosity between corporate management and some of their professionals and the scientists and engineers in the business units. To counter these complaints a number of actions were taken by the corporate lab management.

For example to answer the assertion that nothing was coming out of the labs, corporate management established the red and black book practice. I'm not sure which color represented which, but one was to document technologies that were discovered in corporate and offered to the businesses. Generally speaking, these offerings were made by corporate management going to the business management and insisting that they take a look at the new technology. Almost inevitably the returned without and commitment, and simply told the corporate organization that the business organization was shortsighted. The other book was a list of technologies that had been

transferred, and the commercial successes that had transpired. Everyone in corporate hated the annual call for new entries, and few in the businesses ever knew the books existed. The books were dragged out at the annual show-and-tell for the Board of Directors. Eventually their use waned because they were contributing no real understanding to the problem.

In the late 80's I studied these books and concluded there was a lot of great accomplishments listed that were real and valuable to the company. But, they were pretty much ad hoc, and certainly did not look like the best that could have been done with such an array of great people. I even developed some concepts for improvement by discerning that the corporate labs seemed to do their best work when asked for help by the businesses and when the businesses involved them in the reasoning for the need. I'll get back to this strange conclusion shortly.

Publications and patents were another measure used to prove the value of the corporate investment. The number of patents per million dollars spent of R&D was by far the best in the company, and as we subsequently learned much better than our competition. Our patents were often seminal discoveries and attracted a tremendous level of citation. But, nobody seemed to care about all these metrics. Post docs and professors continued to apply to work in our labs, but

here again the business people were indifferent.

There were many other attempts by both corporate and business management to improve the perceived value of the corporate labs. Once, during the 70's it was decided to split the labs into two organizations. One would continue to do basic science in the way they had done for decades, and the other would try to commercialize the new findings. Within weeks, the basic people stopped talking to the newly ordained applied folks. The latter were viewed as somehow instantly inferior. This experiment lasted a few years and collapsed. There were some attempted business starts, but for the most part they were incompetent and horrendously costly. Without any real business expertise, the amateurs in the corporate labs learned the hard way how tough it is to introduce a new product into the marketplace.

How did the corporate labs get this way? That's an interesting question. If one looks at their history, one can see that they all evolved from the view that in the early days of their companies, it was necessary to get more science into what they made. In the case of DuPont, the need for science was articulated most effectively by an Executive v.p. named Charles Stine in the mid 1920's.

This led to the founding of the corporate labs in

1926, and the hiring of some great scientists and engineers including Wallace Carothers. Within two years of the establishment of the labs, synthetic rubber and Nylon were discovered and on their way to commercialization. The corporate science and engineering labs were deeply involved in all aspects of this and contributed mightily. Carothers himself did not like the discipline of commercialization but was enormously creative nevertheless, because he was right in the middle of all the needs, and couldn't help but be intrigued.

From that spectacular beginning, DuPont launched the polymer era and many new businesses were created. Business units were formed. Each set-up its own R&D, and did extremely good research. The wrote the books on polymer theory, production plant engineering design principles, hazardous materials safe handling and processing and so on. The business units also told the corporate research labs to keep out of their area and do basic science. In effect, go and find some more Nylons, if possible, but for certain keep out of our hair.

So, while initially corporate research was extremely effective because of the quality of the people and the total immersion in the business needs, it grew ever more isolated, and lost its mission. It became obsessed with new and exotic science and less and less

tried to select areas for research based on possible company needs. Rather, where was the most interesting new science that would produce a lot of publications? As you would expect, there were many collaborations with scientists in the business units, but these were ad hoc and only existed because the collaborators liked to work together. Little or no business information and needs were transmitted by these relationships.

In the late eighties, almost all companies that had corporate labs decided something had to be done to make the corporate labs more valuable to the company. Some, in frustration eliminated them. There were many in DuPont that would have been glad to see that course.

It was hard to find a business unit head that thought corporate contributed value to the company. The R&D function heads in the businesses were little different in their opinions of value. Fortunately, several members of the DuPont Executive committee had different ideas. The saw that the problem was not the labs, but the context in which we operated the labs.

They saw that the labs were too insulated from the company, even though they were headed by excellent scientists. Unfortunately these same corporate lab leaders had no knowledge of business, and in fact were

taught to scorn business. And, because of the long history of benign animosity were ignored by the businesses.

A number of us were involved in trying to decide how to approach the problem. We never questioned the quality of the research being done. A number of us were also very strong on the view that telling the corporate organization what to do and what to drop if anything was not the way to proceed. That had been suggested by many of the business R&D management who themselves were not versed in how R&D organizations actually function.

Some companies decided the way to get the corporate organization more integrated with the business needs of the corporation was to change the funding mechanisms. The idea was to make the corporate lab sell their programs to the businesses. At DuPont, corporate science research was historically 90%, and corporate engineering research 25% as a tax on earnings. The latter used to be more than 50% but concern that it was not business driven led to the change twenty years ago. But, by the 1990's it was apparent to many of us that the engineering people were now more interested in selling programs to the businesses than in looking for real breakthroughs. They were on their way to becoming mediocre.

We clearly saw that the problem lay in the isolation of the corporate labs. We also realized that the fix had to be significant and based on a thorough understanding of the needs of the research community.

We approached the isolation problem at multiple levels. Many things were done simultaneously. Here are the major actions we took over a three to five year period.

The head of the corporate research department was moved upstairs, and given an operating director of the labs who was a business R&D v.p. Initially this was met with concern in the corporate labs and by their supporters from the university community. However, this concern melted fast when it was observed how helpful the new laboratory leadership was in creating interfaces with the businesses R&D organizations.

We then took advantage of this new relationship and worked with the business R&D v.p.'s to see where we might work better together without destroying each others capabilities. This quickly led to formation of corporate centers of excellence. These were groups formed in the corporate labs that focused on important science and engineering areas by having the best facilities and an adequate critical mass of outstanding personnel. Professionals from the business R&D labs were often transferred to the

corporate centers of excellence and had the opportunity to work side by side with the basic scientists and engineers in their areas of new knowledge needs. They got to know and respect each other. The centers were managed by corporate, but sometimes the management was imported from the businesses. Often this was for a prescribed period only, but the infusion of new thinking was dramatic in effect. These centers were things like computer science, microwave engineering, catalysis, fluoro chemistry, enzymology basic polymer science and characterization and so on.

By this time the R&D heads from the businesses and corporate had formed themselves into a Corporate Technology Council, and was working together to make the integration even stronger. The principal job of this council which meets monthly, is to get more than the sum of the parts from the companies total R&D organizations. They drive issues like computer standards, setting up of cooperative centers of excellence, and technology networks, adoption of best practices to improve R&D, and even shepherd the development of technology management personnel.

The next major thing done was for the businesses to commit to having the management and appropriate professionals from corporate be an integral part of their business R&D planning. This meant that

corporate people were shown the most intimate needs of the businesses. This had a tremendous effect. Not only were the corporate labs management now aware of the tough technology needs of the businesses, but they often also contributed freely to the ideas of how to solve current major problems. Naturally when they ran their groups in the corporate labs they were much more able to talk about these interesting and challenging problems.

This effect rapidly led to the formation of networks of technology specialists throughout the company. The e-mail system became the key support structure for this activity. Today DuPont has more than 150 such networks in just about every science and engineering discipline you can imagine. A scientist or engineer in DuPont has no hesitancy in consulting people all over the company for ideas and help. The ability to find the best person in the company is only a few keystrokes away. Ten years ago such scientists and engineers lived largely in isolation within there own business units and sometimes even within their own labs at plant sites. Consulting with one another was a rare practice. Today DuPont's more than fifty labs all over the world are in constant communication via individuals using the network directories.

As we evolved the management of the corporate labs we also made a deliberate attempt to swap management with

the business units, and to bring in business astute technology managers from the businesses to lead the major units of the corporate labs.

As a metric of success in this transformation, I can cite the great turnaround in attitude towards the corporate labs by the business management and the Chairman. Today the corporate labs are judged vital in DuPont. They have in fact grown significantly in the last several years, as more activities were consolidated into centers of excellence that corporate was asked to oversee for the company. This, during tough downsizing of almost all organizations within the company. In addition, both the science and engineering labs are 100% corporate supported. That's a customer satisfaction metric if there ever was one.

Other companies have done things to improve their labs. Some appear as good ideas to me, some are a bit naive. But one constant in all these effort are a series of actions to better integrate the corporate labs into the information flow that helps a corporation better link all its elements with the overall mission. No one in the business units is telling the corporate scientists or engineers what to do. They are simply exposing them effectively to planning and strategy sessions devoted to the future needs of the businesses. They do this in all kinds of formal and informal ways, many of which I have just

mentioned. However, the management across the company is driving this culture change.

There are many other example I could list, but I think you get the drift and what I mean when I say that an important element of evaluation for a basic research organization is to examine how it is linked with the total information stream that is part and parcel of the organization and mission it is expected to serve. It's also clear that we need to start with a clear statement of the mission of basic research before we can start to design the metrics to evaluate its context.

In DOE we have been charged to do a better job with this linkage. Many critics have told us they do not believe we are doing a good job in the integration process. These include, the recent Galvin and Yergin studies, the GAO and several studies before these. We cannot continue to do business in the same way as the past and hope that this will all go away. I was in industry for over thirty six years and saw many things come and go. The recent changes responded to a transition that we all recognized was different from the earlier transient ones. I think it's the same with the need that this conference is supposed to address.

And the DOE is taking actions. One of the most recent

is the formation of the Energy Cluster. This is made up of the heads of Fossil Energy, Energy Efficiency and Energy Research. No real actions have yet come from this combination, but it is a powerful group who could make real progress in one of our major mission with respect to more effective integration. A second major initiative under way is the collaboration of Energy Research and Environmental Management to plan research to improve the basic understanding of environmental cleanup problems. And finally, we have formed a Research Council that will work across all the Program areas in DOE to achieve better integration between basic research and the applied programs. I have been asked to chair this group, which is composed of all the appropriate Program heads and their chief deputies. This is a very similar group to one I headed in DuPont. In DuPont the idea was to integrate the corporate labs more effectively with the business needs of the company. I have a lot of enthusiasm for this new structure.

This gets me back to today's subject, the evaluation of basic research. I think the integration activities I have just mentioned are key to predicting the success of research. It is important to do case studies, develop and use more innovative econometric models, and strive for ways to measure impact. Our critics are crying for this. But, as I tried to illustrate by my remarks, I think the context of

research is really what matters most. If the context is poor the research will be inefficient, ad hoc, and at times downright poor. All the peer reviews in the world can not overcome a poor context.

Consequently, I hope you have and will continue to spend at least some of your thinking time trying to frame the evaluation questions and metrics that look at the context of research here in government. It seems to me we have not spent enough time doing this as an intellectual community. We keep promising numbers, but we don't deliver numbers that anyone believes or cares about. Perhaps this is telling us something. In the DOE we are trying to develop more mechanism that involve the basic and applied organizations in genuine information exchange in the context of the DOE's mission. Simply put, if we understand the mission, then we ought to be able to devise context improving mechanisms that are appropriate to improve the effectiveness. These should also be measurable but they will not look like what the accountants and other critics expect.

We who understand research will have to educate them rather continuing to pretend we can give them numbers that predict the value of basic research. They are in a real way like customers of a company. They think they know what they want, but if you simply give them what they want without innovation, you will disappoint

them time and time again.

Some thoughts in closing that come to my mind that you may wish to think about to help evaluate the health of the context for basic research in DOE are:

Does the mission of basic research usefully prescribe parameters of appropriate scope?

How can we evaluate the degree and quality of co-planning that is actually done in the DOE between the basic and applied programs? What are the best mechanisms to do this and to track improvement?

How much mutually funded collaborations should the applied and basic programs do to achieve optimum integration?

Does the DOE have networks of professionals that self organize and help each other? If not, what are the barriers?

What kind of relationships do the management of the applied and basic programs have with each other? How can this be measured and improved?

How knowledgeable are the basic and applied

> program management about each others needs? My recent interviews with most of them convinces me we are not in best possible shape. Certainly, within a modern company, the status would be deemed unsatisfactory.

I recognize that many will say these are all input measures, but I guarantee that they are also critical indicators of the value being produced. If these things are done well and continuously improved, the other measures you develop and follow will also be much improved. Good companies know this. What makes us so different?

PRELIMINARY AGENDA -- R&D EVALUATION WORKSHOP
DOE OFFICE OF ENERGY RESEARCH
SEPTEMBER 7 AND 8, 1995

DAY 1

8:30 - 9:30 Opening Remarks with Questions and Answers -- Martha Krebs, Director, Office of Energy Research

9:30 - 10:30 Panel presentation of example case -- Intermetallic Alloys -- DOE research with impact on DOE mission, applied research, industry and society -- Iran Thomas (Office of Basic Energy Sciences), Linda Horton and others (Oak Ridge National Laboratories)

10:30 - 10:45 Break

10:45 - 11:30 Panel discussion on current ER evaluation efforts -- John Moore (Chairman, Panel on Value of Basic Research) and others

11:30 - 12:00 Discussion on workshop approach

12:00 - 1:30 Lunch
Speaker: Erich Bloch, Distinguished Fellow, Council on Competitiveness

1:30 - 3:30 Breakout Sessions -- expert discussion of assigned questions
- Session 1: Case studies and TRACES approach
- Session 2: User surveys and ROI/Econometrics
- Session 3: Citation analysis and Expert panels

1:30 - 2:45 Press Briefing Roundtable, M. Krebs, I. Thomas, A. MacLachlan, J. Moore

3:45 - 5:00 Integrated Session - expert discussion of emerging themes

DAY 2

8:30 - 9:30 Perspective of cooperative research development -- Al MacLachlan, Deputy Undersecretary for R&D Management

9:30 - 11:30 Breakout Sessions - expert discussion of assigned questions

11:30 - 12:30 Summarize preliminary findings of workshop

12:30 - 1:30 Lunch

1:30 - 4:00 Experts discuss questions, road map for future research

4:00 PM Adjourn

U.S. Department of Energy
Office of Energy Research (ER)
R&D Evaluation Workshop

Date and location: Thursday and Friday, September 7 and 8
Holiday Inn Capitol, 550 C Street SW, Washington, DC

Objective: To promote discussion between experts and research managers on assessing the impact of DOE basic energy research upon the energy mission, applied research, technology transfer, the economy, and society. The purpose of this impact assessment is to demonstrate results and improve ER research programs in this era when basic research is expected to meet changing national economic and social goals. Experts will discuss innovative methods as they could be applied to an example case: research in intermetallic alloys, funded by the Office of Basic Energy Sciences (OBES) beginning in the 1970's.

Specific questions to be addressed include:

1. By what criteria and metrics, considering internal and public uses of the information, can OBES measure performance and evaluate its impact on the DOE mission and society *while maintaining an environment that fosters basic research?*
2. What combination of evaluation methods, including innovative techniques, best applies to assessing the performance and impact of OBES basic research? The focus will be upon these methods:
   - Case studies
   - Citation analysis
   - User surveys
   - Return on DOE investment, econometrics
   - TRACES approach
   - Expert panels
3. What combination of methods and specific rules of thumb can be applied to capture impacts along the spectrum from basic research to products and societal impacts?

A tentative agenda is attached.

Special Invited Guests:
Staff Directors of House and Senate Committees on Science and Technology
Representatives of Office of Science and Technology Policy, Office of Management and Budget
Representatives of the National Academy of Sciences
Representatives of other Federal agencies

Participants in Expert Discussion

Research Managers Invited:

- ER Associate and Division Directors
- Representatives of DOE laboratories and applied research and technology partnerships offices
- Members of the Panel on the Value of Basic Research
- Project Managers and Technology Steering Group of the OBES Center of Excellence for the Synthesis and Processing of Advanced Materials

R&D evaluation experts to be invited: (* indicates confirmed attendance as of 8/3/95)

- Francis Narin*
- Susan Cozzens
- Len Lederman*
- Keith Pavitt
- Harvey Averch*
- Edwin Mansfield
- Albert Link*
- George Teather*
- Zvi Griliches
- Anthony F.J. Van Raan
- Barry Bozeman*
- J. David Roessner*
- Ronald Kostoff
- Julia Melkers
- Maria Papadakis
- Daryl Chubin

8/22/95

PRELIMINARY DRAFT OUTLINE -- Basic Energy Sciences R&D Evaluation

I. Purpose of study

II. The Role of Evaluation
- A. Very brief survey of literature on evaluation techniques for federal research
- B. Multiple purposes of evaluation
  1. Program improvement
  2. Public support
  3. Meeting requirements (GPRA, OMB)

III. DOE research goals and objectives
- A. Goals
- B. Criteria for success
- C. Environment for innovative research

IV. Current practices in evaluation within DOE
- A. Program manager review
- B. Peer review, expert and advisory panels
- C. Performance measures (contract reform, pilot study)
- D. Case studies (OBES)
- E. Citation analysis, Surveys
- F. Return on Investment (Sandia study)
- G. Do we satisfy requirements?

V. What would comprise an ideal evaluation system?
- A. What decisions will be made possible or be improved, and how will this effort make R&D more effective?
- B. What are the right questions to ask? What performance measures and evaluation questions are priority?
- C. What innovative combinations of methods can be applied, at what frequency, and at what level of program?
- D. Are there conflicts between good practice and mandates?

VI. Suggestions for action
- A. Rules of thumb for moving toward an ideal system
- B. Advantages and disadvantages of these rules. Minority views will be recorded.

Today, well over 50 companies incorporate features of these new techniques in their own ceramic processing facilities. The sizes of their in-house programs on the new processing techniques range, in some cases, to more than 100 people. Published papers and reports indicate that foreign firms are also applying these concepts.

Ceramics fabricated by these methods and their extensions are now being used in a variety of applications. These include automobile turbochargers (rotors), microelectronic circuits, high-temperature energy converters, engine parts, computers, televisions, aircraft, and vital defense weapons, among others.

More generally, many varied ceramic compositions with fine grains can now be fabricated. Major improvements in the mechanical and electromagnetic properties of ceramic formulations have been reported as a result of better processing. Previously unrecognized avenues of materials fabrication, such as layering of different ceramic materials, have been opened up, and totally new electronic components and devices are now possible as a result.

The original decision to eschew further tinkering with recipes and to proceed, instead, with more fundamentally oriented research into the chemical processing of ceramics led directly to these results. In this way, Basic Energy Sciences provided the seed money and critical initial sponsorship of a risky venture. This work, in turn, generated the concepts and the experimental verification that allowed others to appreciate the advantages of the new approach and encouraged them to pursue practical applications.

NICKEL ALUMINIDE

The continuing quest for improved efficiency in the use of combustible fuels is often limited by the properties of the materials used to make up the fuel burning engines. In theory, higher operating temperatures make possible higher efficiencies. In practice, the physical limitations of materials call for moderation to ensure engine reliability and longevity.

Recently, Basic Energy Sciences researchers contributed significantly to the emergence of a new metallurgical alloy, called nickel aluminide, which promised to combine the best of both worlds. It is made up of nickel and aluminum, combined in a ratio of about 3 parts to 1, with trace amounts of boron added. When properly prepared, it exhibits extraordinary properties.

For example, nickel aluminide is much stronger than steel. In contrast to most other materials, it actually increases in strength with higher temperatures, up to about 1300°F. It then maintains this strength to more than 1600°F. It is also lightweight and strongly resists corrosion by oxidation.

Important for fabrication and durability, it is malleable and ductile. This makes the material easy to form into different shapes and forgiving to shock and stress. Yet, with the addition of certain other elements, it can be made to resist strongly permanent deformation and rupture at high temperatures—two common modes of metal failure.

The mechanical properties of nickel aluminide provide significant advantages over many currently available heat resistant materials in a number of important applications. These range from machine tools and boilers to parts for the automotive and aerospace industries. For example, at 1500°F, a temperature well within the operating temperature ranges experienced by most metal parts in today's gas turbines, jet engines, and diesels, nickel aluminide is four times stronger than most high-temperature speciality steels.

Because of its high strength and resistance to oxidation at these temperatures, the alloy is now being examined under exclusive license by Cummins, Inc., a major United States manufacturer of heavy diesel engines. While more expensive than most ordinary steels, nickel aluminide promises to be cost competitive with other specialty materials, such as heat resistant alloys made of nickel, titanium, chromium, and cobalt.

An initial application under investigation is its use in strengthening exhaust valves. This part of the engine, which is subjected to high temperatures, corrosive environments, and repeated pounding, is often the first part to fail. Extending valve life with an improved material would increase the engine's reliability, stretch out maintenance schedules, reduce costs of repairs under warranty, and make the entire engine more competitive in international trade. Ultimately, new designs may capitalize on the added strength, raise combustion temperatures, and improve fuel efficiency.

A more recent area of emphasis is its use in strengthening the rotor blades in so-called "turbocharger" gas compressors. Propelled by the hot exhaust gases of combustion engines, turbochargers compress intake air and boost significantly overall engine efficiencies. Nickel aluminide not only stands up to the extreme heat of the exhaust gases, but also exhibits a much longer life before failing due to cyclic stresses and fatigue. It also costs less than competing "superalloys" and other high-temperature withstanding materials.

Because the alloy is made partly of aluminum, it is 10 percent lighter than steel. Also, because of its strength, a part made from the alloy can be designed smaller, further reducing its weight. Hence, the alloy also has potential applications in the aerospace industry as a substitute for heavier materials now required for strength, such as fasteners, rivets, and certain structural components.

Finally, another valuable property of the alloy is that once it has formed an initial, protective layer of aluminum oxide, it is nearly impenetrable to further corrosion and oxidation. Its corrosion resistance has been measured at 1,000 times better than competing steels. As a result, it is now being investigated for use in heat processing equipment subjected to fouled environments, such as steam boiler tubes and hot exhaust gas heat recovery equipment in industry.

Technically, nickel aluminide is part of a larger family of materials known as "ordered" intermetallic alloys, so named because of the precise ordering and interweaving of the atomic structures or lattices of the two metals. This particular situation, called a low "free energy" condition, makes it difficult to remove an atom from its position in the lattice, which gives the material its strength and chemical stability.

Intermetallic alloys were well known in the 1950's and 1960's for their extraordinary strength. Unfortunately, the problem with them in the past had always been that they were too brittle for most practical applications. Although single crystals of some of these alloys were known to be ductile, bulk quantities in polycrystalline form fractured, like glass.

The basic problem was with the microscopic interfaces, called grain boundaries, where the crystals which constitute the bulk material join together. It is at these interfaces where the crystals tended to pull apart or slip against each other under stress, causing the material to break.

This problem could be solved, it was hypothesized, if means were found for increasing the adhesiveness of these surfaces. Perhaps the addition of small amounts of alloying elements, to be used as "impurities" in the larger matrix of nickel and aluminum, could somehow cause a strengthening of metal-to-metal bonding at these surfaces. Little knowledge was available to guide this search, however, and progress in solving the brittleness problem slowed.

In the late 1970's, Basic Energy Sciences researchers were working on a seemingly unrelated problem. In retrospect, this work laid a foundation of basic knowledge and improved laboratory capabilities in the use of a key instrument. Both investments later helped to explain a startling new discovery and set into motion a resurgence of research activity on ordered intermetallic alloys.

These researchers were studying the effects of neutron radiation damage on certain types of stainless steels used in nuclear reactors. They would stress the steel until it would break. Then they would study the chemical and atomic composition of the fractured surfaces.

They found that, while the steel had remained in solid polycrystalline form, certain alloying elements, such as phosphorus, tended to migrate to the intergranular fracture surfaces and concentrate there in amounts much larger than normal bulk proportions would predict. Further, they determined that other alloying elements, such as carbon, oxygen, and chromium, did not exhibit this migrating phenomenon, suggesting atomic selectivity.

Importantly, the researchers increased their capabilities in the use of one special research tool called Auger electron spectroscopy. This instrument bombards the top few layers of atoms on a given surface with low energy electrons. This temporarily disturbs the equilibrium nature of the electron shells of these atoms, ultimately resulting in the emission of an identifying spectrum of electrons, whose abundance and energies can be detected and measured. From this spectrum, the constituent elements of the top-most surface layers of a material, including their relative atomic proportions, can be accurately inferred.

Further, the instrument can "raster" the observed surface with argon ions, blasting off the top few layers of atoms like a machine gun. This then allows subsequent analyses of the underlying layers. Through repeated cycles of this process, comparative analyses of one layer after another can be made. This reveals the depth of impurity segregation on the intergranular surfaces and portrays a good picture of surface chemistry and its effects on bulk material properties.

In 1979, Japanese researchers reported a remarkable discovery. They found that the addition of small amounts of boron to nickel aluminide increased its ductility. Following this lead, researchers at Oak Ridge National Laboratory, using Exploratory R&D funds, determined that this phenomenon only worked under a highly specific condition. This was when the total number of aluminum atoms, compared to the total number of nickel atoms, was just slightly less than that dictated by its natural or stoichiometric ratio. In nickel aluminide (Ni_3Al), the natural atom ratio of nickel to aluminum is 3 to 1, or in other terms, 75 to 25 atom percents.

In 1982, Basic Energy Sciences researchers applied their earlier gained knowledge about grain boundaries, intergranular surfaces, and impurity migration to the problem of understanding what was going on. Once again, a key element of their research involved the capabilities of Auger electron spectroscopy.

In a report published in 1985, Basic Energy Sciences researchers presented their findings and offered an explanation of the ductility phenomenon. Under conditions where the relative abundance of aluminum, compared to nickel, was slightly less than the natural ratio in pure Ni_3Al, say 24 atom percent rather than 25, boron atoms migrated in droves to the grain boundaries. They accumulated there in the top two or three atom layers of the surface. They concentrated themselves in numbers far outweighing, by 60 times or more, their bulk proportion of, say, 0.1 percent.

Under these circumstances, boron acts as an electron donor in the lattice structure. This is believed to add to the electron bonding potentials of nickel atoms between the intergranular surfaces. This makes the surfaces adhesive, lending ductility to the polycrystalline bulk material.

Other impurities, by contrast, such as sulfur or phosphorous, were found to migrate strongly to open cavities and voids, and to a much lesser degree to the grain boundaries. This was fortunate, because these elements act as electron captors, believed to diminish the bonding strength between the intergranular surfaces, encouraging fracture and adding to embrittlement.

Specifically, Basic Energy Sciences researchers showed that the solubility limit of boron in Ni_3Al was about 0.3 weight percent; that ductility of Ni_3Al increased dramatically from near zero to over 50 percent elongation with the addition of boron up to about 0.1 weight percent; that boron migration to the grain boundaries was highly dependent on the existence of slight deficiencies in the relative abundances of aluminum atoms compared to nickel; that grain boundary boron segregation strongly affected grain boundary cohesion and related atomic arrangements, which affected ductility; and that distribution of other impurities remained unaffected by the existence of slight variances in alloy stoichiometry (relative atom abundances).

All of this led to a much clearer understanding of the boron-ductility phenomena which, in turn, led to more broadly-based research on other members of the family of ordered intermetallic alloys. It also provided much needed specificity to guide theoretical work on the role of atomic arrangements and electron structures in metal-to-metal bonding at grain boundaries. Finally, it encouraged engineers to pursue practical applications by lending predictability to various metallurgical procedures.

The discovery of the boron-ductility effect, accompanied by this detailed understanding, was pivotal in the development of a whole new field of research. It precipitated much follow-on and continuing research by both Government and industry on nickel aluminide. As one measure of industry's interest, over a half a million dollars in research money was provided by private companies to the Oak Ridge National Laboratory in 1985 to investigate related production and processing methods. The research earned an IR-100 Award from *Research & Development* magazine (formerly *Industrial Research*), and the alloy is now subject to extensive patent and licensing activity by industrial firms throughout the United States.

Oak Ridge Report

INDUSTRIAL TECHNOLOGIES

America's industries consumed 39% of the nation's end-use energy in 1990. A small number of major manufacturing groups (primary metals, petroleum refining, chemicals, pulp and paper) account for about 70% of this industrial energy use, or about 27% of the nation's total energy use. The DOE/EE Office of Industrial Technologies (OIT) has begun an initiative to work with these industries to cut their nonproductive energy use and environmental costs. Called "Industries of the Future," the initiative is targeted at the steel, aluminum, foundry, petroleum refining, chemicals production, pulp and paper, and glass industries. The research programs will be developed by the industries and will be jointly executed by industry, universities, and national laboratories.

ORNL has made technological advances that are contributing to improved industrial efficiency through decreased energy consumption, improved product quality, reduced equipment downtime, and decreased waste streams. As industry expenditures on nonproductive costs decrease, resources are made available for market expansion and investment in plant and capital equipment. ORNL's goal in research for the DOE/EE Office of Industrial Technologies is to assist U.S. industry in capturing and maintaining global market share through technological improvements.

ADVANCED MATERIALS AND MANUFACTURING TECHNOLOGIES

Advanced Industrial Materials. The Advanced Industrial Materials (AIM) Program develops new and improved materials and manufacturing technologies leading to more efficient use of energy in support of the Industries of the Future initiative in the Office of Industrial Technologies. High-temperature intermetallic and metallic alloys with high ductility, corrosion resistance, and strength are being developed. Metal-bonded composites are being developed for optimal use at temperatures between those of currently available alloys and ceramics. Coatings and engineered porous materials are being evaluated for various applications. Microwave technology is being pursued because it offers new and exciting potential for materials with unique properties. Recently, ORNL has participated in an activity focused on performing materials needs and opportunities assessments for the pulp and paper industry. This effort–which includes participants from industry, universities, institutes, other national laboratories, and DOE–has resulted in projects aimed at developing improved materials to meet the identified needs of the pulp and paper industry.

One of the most successful ORNL inventions resulting from the AIM Program is ordered intermetallic alloys, which were developed in a cooperative project with the DOE Office of Basic Energy Sciences. Ordered intermetallic alloys are different in atomic structure from conventional alloys, and the atomic structure can be optimized for specific applications. One of the most successful of the ordered intermetallic alloys is nickel aluminide. Seven industrial firms have held exclusive and nonexclusive licenses to the material, and optimization for commercial applications is the subject of several CRADAs (see the highlight "Nickel Aluminide Alloy Finds Commercial Application in the Steel and Die-Making Industries").

A project at the Georgia Institute of Technology has successfully demonstrated the feasibility of producing monosize hollow spheres of many ceramic compositions on a production basis. The mechanical strength and thermal conductivity of the spheres have been documented, and mathematical modeling of the sphere-forming process has been successful. The technology, trademarked

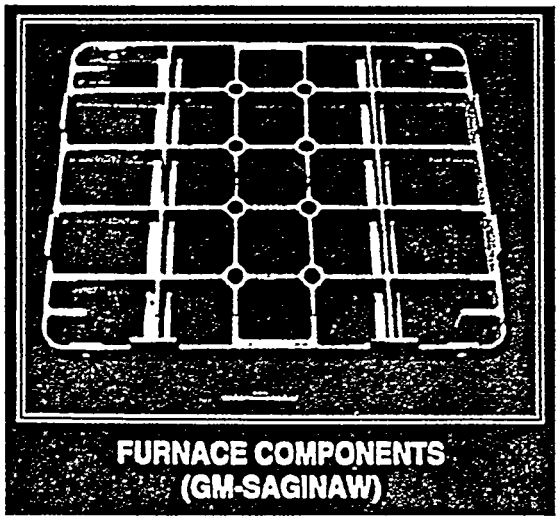
NICKEL ALUMINIDE ALLOY FINDS COMMERCIAL APPLICATIONS IN THE STEEL AND DIE-MAKING INDUSTRIES

In a CRADA with a small business called Metallamics, nickel aluminides are being developed as roller material in heat-treating furnaces used in the production of plate steel. The objective is to produce a roll material that will prevent scratching of the plates as they move over the rolls and offer longer service life than current materials. Field trials in a steel mill furnace are being conducted with promising results: Metallamics expects to fully commercialize the technology by June 1996. ORNL is aware of at least four steel companies that are waiting to install the rolls when they are available. The companies expect to recoup their investment in nickel aluminide rolls through decreased furnace downtime, improved product quality, and energy savings.

In a CRADA with Rapid Technologies, Bimac, and the DOE Office of Basic Energy Sciences, nickel aluminides are being made into walking-beam furnace rails. The rails must resist oxidation and have good high-temperature strength. Commercialization is expected by the end of 1995. Bimac and Rapid Technologies expect that the increased operating temperatures possible in furnaces through use of nickel aluminides will save 1 trillion Btu of energy over 2 years, with attendant decreases in carbon dioxide emissions.

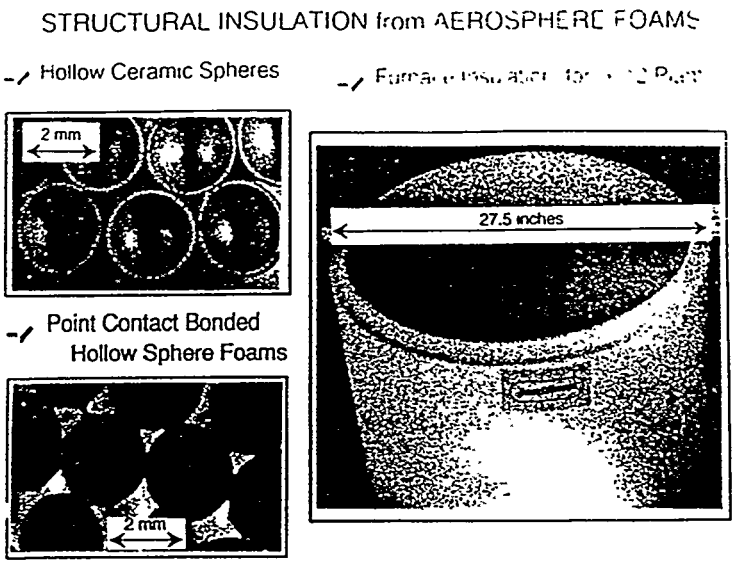
Fixtures for use in carburizing heat-treating furnaces are being made from nickel aluminides. A CRADA with General Motors focuses on fixture manufacture and testing under production conditions, and commercialization is expected by June 1996. If the trial results continue to be encouraging, General Motors plans to completely replace its current furnace fixture material with nickel aluminide. In addition, several other companies have expressed interest in the application.

A more unusual application for nickel aluminide is as a die material for rare-earth magnets. Magnet disks are hot-pressed from the dies; nickel aluminide is the only material known to possess the required high-temperature strength and to be compatible with the rare-earth compound. This application has been fully commercialized in cooperation with General Motors and Metallamics (the rare earth compound, NdFeB, was developed at General Motors). The dies are part of a license that General Motors offers to other companies, and they have been licensed to companies in Germany.



Nickel aluminides offer savings for high-temperature applications, such as furnace fixtures (left) and rollers (right)

Aerospheres™ makes it possible to form hollow spheres from dispersions of inexpensive ceramic powders, engineer the sphere wall to minimize heat conduction, and bond the spheres into structural monoliths, such as the example shown here. Applications include structural insulation, radiant gas burners, liquid metal filters, low mass kiln furniture, and particulate filters. Significant discussions are ongoing with the Gas Research Institute in the area of radiant burners and with an electronics company for kiln furniture. In addition, tests with a major electronic equipment supplier have shown that aerosphere insulation works well in the vacuum/corrosive gas environment used for thin film continuous vapor deposition.



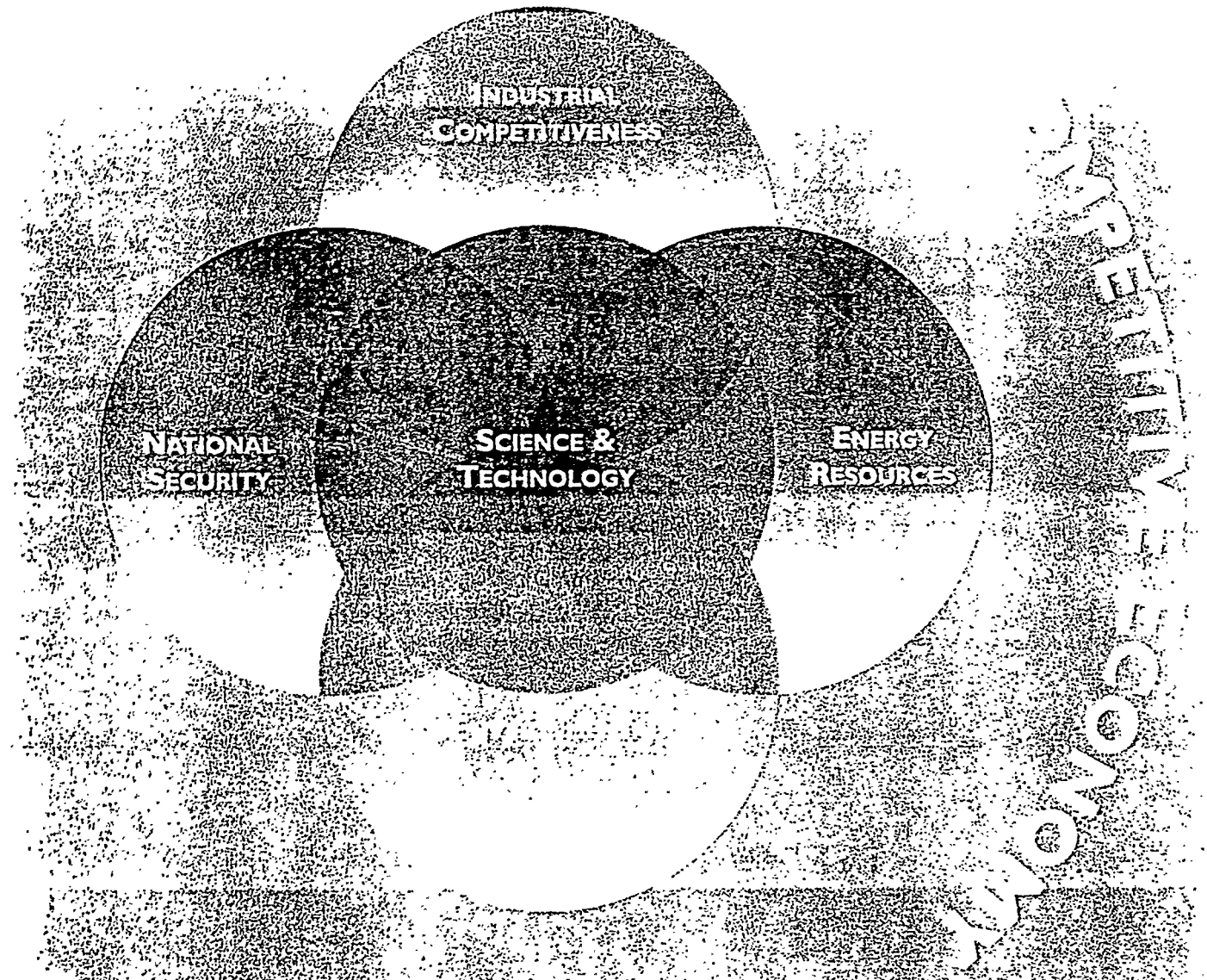
Structural insulation from Aerosphere™ foams.

Continuous Fiber Ceramic Composites. The Continuous Fiber Ceramic Composite (CFCC) program focuses on development of processing methods for fabrication of CFCC components for industrial applications. CFCCs offer high-temperature stability, corrosion resistance, and light weight; and they have the potential for providing significant energy and environmental benefit to U.S. industry. High-efficiency, high-temperature heat exchangers and gas turbines made with CFCCs could save U.S. industry as much as $2 billion per year in energy costs. Annual nitrous oxide emissions from industries could be cut as much as 917,000 tons, and annual carbon dioxide emissions by 118 million tons.

The CFCC development program is being implemented through joint projects with industry. ORNL provides technical assistance in project evaluation and industry interaction, as well as an R&D program that provides improved technologies for the design, development, fabrication, and characterization of composite materials to meet different needs, including the development of methods for testing these new materials. An example of one of these joint projects is provided in the highlight "Advanced Ceramic Composite Material for Industrial Gas Turbine Combustor Liners–The GE Team's Approach."

Advanced Bioprocessing Concepts. ORNL investigates the separation and processing of industrial chemicals using advanced bioprocessing concepts. The focus is on innovative bioprocesses that can exploit renewable resources such as corn sugars, woody biomass, or even carbon dioxide or water, for the production of fuels and chemicals. These advanced bioreactors can be used in aqueous, gaseous, and nonaqueous systems. Columnar bioreactors have the potential for significant increases in volumetric productivity, perhaps as much as a tenfold increase over conventional technologies (see highlight "Fluidized-Bed Bioreactor Economics Look Good").

Much of the work on advanced bioprocessing concepts takes place in the Bioprocessing Research and Development Center and employs the equipment of the Bioprocessing Research User Facility described in Chapter 1. The center was established in 1991 to capitalize on ORNL's pioneering



STRATEGIC PLAN

multi-program laboratories, 10 single-purpose laboratories, 11 smaller special-mission laboratories, and a wide range of special user facilities critical to U.S. industry's global competitiveness.

In fiscal year 1994, the Federal Government's total funding for research and development was $72 billion, spread across 24 agencies. The Department of Energy's share of this research, $7 billion, is the fourth largest and represents almost 10 percent of the total Federal spending.

Recent breakthroughs emanating from the Department's system of laboratories include:

- The world's record in photovoltaic energy conversion efficiency at the National Renewable Energy Laboratory.
- The world's record for fusion power levels produced at the Princeton Plasma Physics Laboratory.
- The world's most powerful source of "soft" x-rays at the Lawrence Berkeley Laboratory.

The Department has extended its basic science with a new emphasis on applied research and partnering with industry. This is best exemplified by the Clean Car Initiative, a Cooperative Research and Development Agreement, negotiated with General Motors, Chrysler, and Ford to develop efficient, clean vehicles that are practical and affordable. Other examples of innovative partnerships include DOE defense technology that is now being used to reduce medical radiation doses and provide better images of mammograms, a broad-based partnership with the integrated textile industry (AMTEX), and a new process for soldering printed circuit boards that eliminates the use of ozone-depleting chemicals while saving energy.

We are the leading Federal agency in patent applications with more than 1,000 from 1990 to 1992, as well as the leading agency in licenses granted with more than 400 during that same period. As an example, the Los Alamos National Laboratory developed and patented an acoustic resonant ultrasound spectroscopy technology to detect defects in aircraft wheels and that is now being used to determine the structural integrity of bridges throughout the Nation.

In 1993, the Federal Government received 34 "R&D 100 Awards" given annually for the most important inventions—DOE won 26 of them. An example of an award from 1992 is the solar water detoxification system which has become part of a Cooperative Research and Development Agreement with industry. The system uses sunlight and a nontoxic catalyst to destroy hazardous organic substances in groundwater and industrial waste water.

Our Mission

We possess the human and physical assets to achieve the mission that follows:

The Department of Energy, in partnership with our customers, is entrusted to contribute to the welfare of the Nation by providing the technical information and the scientific and educational foundation for the technology, policy, and institutional leadership necessary to achieve efficiency in energy use, diversity in energy sources, a more productive and competitive economy, improved environmental quality, and a secure national defense.

Vision

By the turn of the century, the Department of Energy through its leadership in science and technology will continue to advance U.S. economic, energy, environmental, and national security by being:

- A key contributor in ensuring that the United States leads the world in developing, applying, and exporting sustainable, clean, and economically competitive energy technologies.
- A key contributor in maintaining U.S. global competitiveness through leadership in environmentally-conscious materials, technologies, and industrial processes.
- A major partner in world class science and technology through its national laboratories, research centers, university research, and its educational and information dissemination programs.
- A world leader in environmental restoration, waste management, and pollution prevention.
- A vital contributor to reducing the global nuclear danger through its national security and nonproliferation activities.
- A safe and rewarding workplace that promotes excellence, nurtures creativity, rewards achievement, and is results-oriented and fun.

Core Values

The Department will succeed only through the efforts of its people. How well we perform individually and collectively is a function of the beliefs and values that motivate our behavior. The employees of the Department of Energy have chosen the following core values to serve as guideposts and our conscience in fulfilling our mission and achieving our vision.

1. We are customer-oriented.
2. People are our most important resource.
3. Creativity and innovation are valued.
4. We are committed to excellence.
5. DOE works as a team and advocates teamwork.
6. We respect the environment.
7. Leadership, empowerment, and accountability are essential.
8. We pursue the highest standards of ethical behavior.

The Total Quality Philosophy

Our core values will define our culture. Our culture will help us achieve our vision to fuel a competitive economy. A philosophy of total quality management and continuous improvement will serve as the foundation to meet the needs of our customers and allow us all to maximize our potential and make work rewarding.

Total quality will be achieved through customer satisfaction, leadership commitment, continuous improvement, labor/management partnering, and employee involvement. Our journey towards total quality has already begun, and there are many important efforts underway that support this new approach. Examples include customer service plans, process improvement teams, leadership training to support our core values, and implementation of total quality guidelines.

Employees and management working together on these key initiatives will empower all of us to improve customer satisfaction, focus our energy on value-added products and services, and make our jobs more rewarding.

DOE's Five Businesses

In response to world changes and today's new challenges and priorities, we took a fresh look at our business lines. What we

found were mission areas that operated in a vacuum from one another. There was little synergy or integration of departmental assets. In general, little communication across organizational lines occurred. We found an organization structured to meet demands and challenges that were no longer relevant. We recognized that our science and technology capabilities had not been strategically leveraged. We decided to fundamentally reorient both the nature of our businesses and how they were managed.

Through our strategic planning efforts, we identified five businesses that most effectively utilize and integrate our unique scientific and technological assets, engineering expertise, and facilities for the benefit of the Nation. These new businesses which directly affect the security and the quality of life of every American, are:

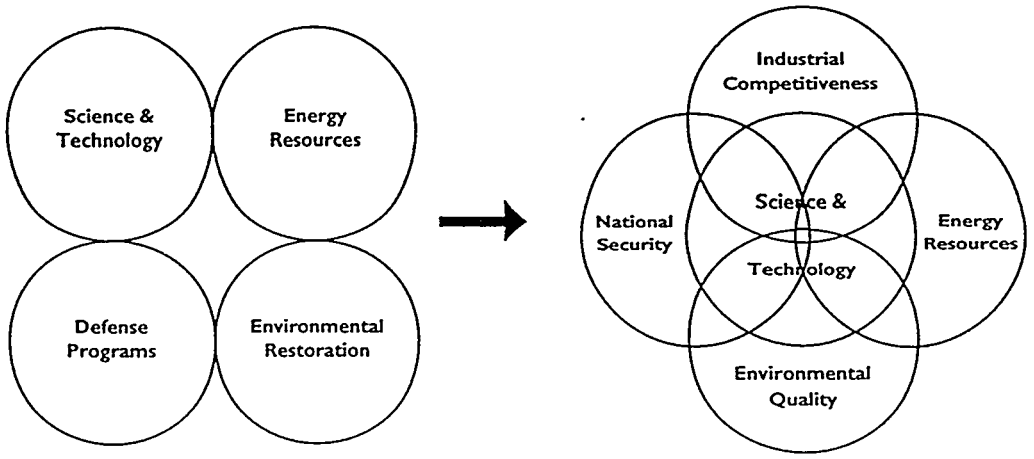
Industrial Competitiveness: Promote economic growth and the creation of high-wage jobs through research and development partnerships with industry, drive products into the domestic and international marketplace, and help industry become more competitive by cost-effectively shifting from waste management to resource efficiency and pollution prevention.

Energy Resources: Encourage efficiency and advance alternative and renewable energy technologies; increase energy choices for all consumers, assure adequate supplies of clean, conventional energy, and reduce U.S. vulnerability to external events.

Science and Technology: Use the unique resources of the Department's laboratories and the country's universities to maintain leadership in basic research, increasingly focus applied research in support of the Department's other business lines, and maintain world technical leadership through long-term, systemic reform of science and mathematics education.

National Security: Effectively support and maintain a safe, secure, and reliable enduring stockpile without nuclear testing, safely dismantle and dispose of excess weapons, and provide the technical leadership for national and global nonproliferation activities.

Environmental Quality: Understand and reduce the environmental, safety, and health risks and threats from DOE facilities and decisions, and develop the technologies and institutions required for solving domestic and global environmental problems.



The Department Has Fundamentally Reoriented Its Business Lines For The Benefit Of The Nation.

- By 1995, determine a long-term waste repository program funding policy and profile, rebaseline the Yucca Mountain site suitability effort, and, by 1996, define the departmental role regarding nuclear spent fuel interim storage at reactor sites and in the Federal waste management system.
- Increase in percentage of U.S. market share in the export of clean energy technologies.

GOAL 4

Promote economic and regional equity for all Americans through changes in the systems of energy production, delivery, and end-use.

STRATEGIES

- Develop tax policies and fund programs that ensure universal access to affordable energy services.
- Develop policies that eliminate disproportionate adverse environmental effects of energy systems on geographic regions, minority and low-income groups, and local communities.

SUCCESS INDICATORS

- Increase in percentage of public utility commission decisions giving explicit recognition to equity issues.
- Increase in number of low-income households weatherized.
- Decrease in ratio of energy system costs to benefits, by population groups.
- Increase in equity considerations in the siting of new energy systems.

SCIENCE AND TECHNOLOGY

First-class basic and applied science are needed to advance industrial competitiveness, clean energy resources, national security, and environmental quality through technology leadership. The Administration's technology plan of February 23, 1993, recognizes this by setting a key goal for the Nation of world leadership in science, mathematics, and engineering.

VISION

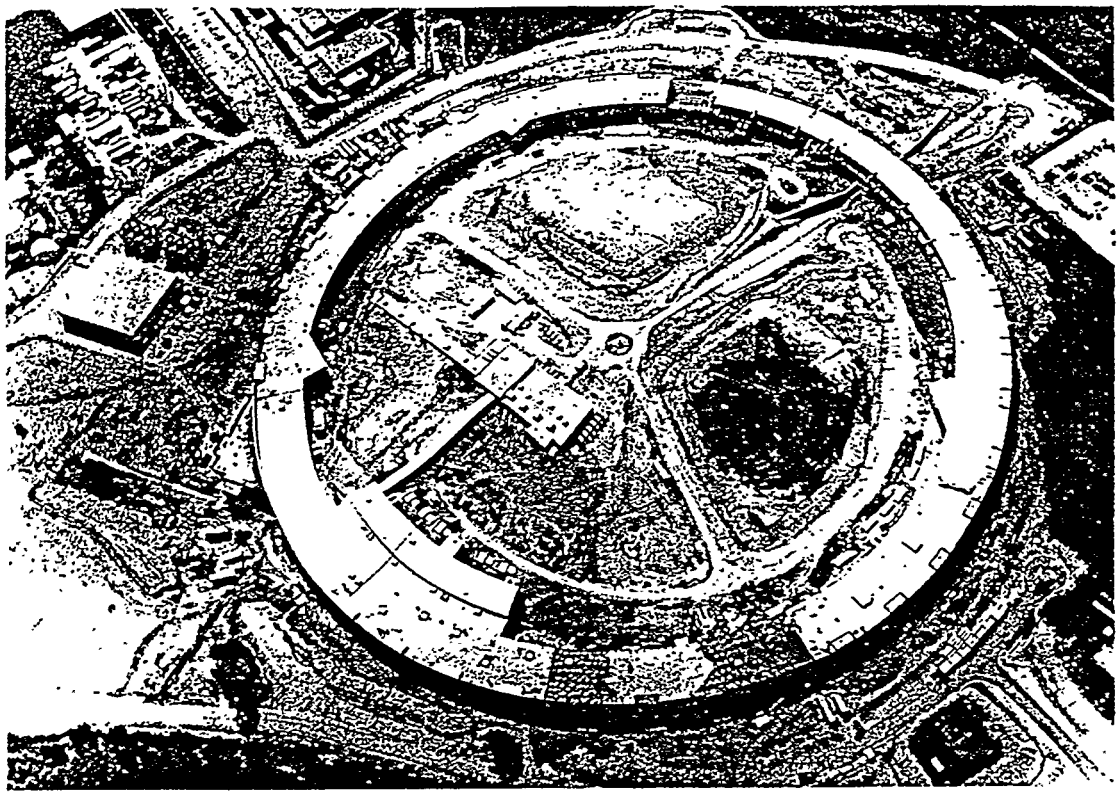
Science and Technology provide the knowledge that drives our future. World-class scientists and engineers; working in world-class facilities on leading-edge problems will spawn the knowledge that revolutionizes technology—the knowledge and technology that others need to achieve their vision.

GOAL 1

Provide the science and technology core competencies that enable DOE's other businesses to succeed in their missions.

STRATEGIES

- Maintain and validate program excellence and balance in basic science and applied science that supports the energy, environment, national security, and industrial competitiveness missions.
- Fully utilize research facilities, as appropriate, to reduce unit costs.
- Develop innovative options for funding R&D partnerships.
- Build on and nurture appropriate DOE core competencies.
- Encourage flexibility in research programs.



DOE's advanced photon source will provide research opportunities that could lead to higher quality products that last longer. When experiments begin in 1996, it will produce x-ray beams one trillion times more brilliant than conventional x-ray machines. Joint research teams from industrial, university, and government labs will build and operate research facilities at the site.

- Improve communications and establish partnerships among suppliers, customers, and stakeholders.

Success Indicators

- Quality of science, as indicated by favorable outside peer reviews and judgments of expert advisory committees.
- Closer linkage of energy research programs to DOE's energy, national security, and environmental technology programs.
- Maintain or improve the performance and preeminence of the Department's large research facilities, as indicated by the reliable and cost-effective operation and maintenance of world-class research facilities and endorsements from the research users.
- Increase in DOE's influence in developing the information superhighway.

Goal 2

Provide new insights into the nature of matter and energy, address challenging problems, and create a climate in which breakthroughs occur.

Strategies

- Maintain and validate program excellence and balance in high energy and nuclear physics and other fundamental sciences.
- Ensure a flow of knowledge into society by striking a reasonable balance in support of principal investigators, new facilities, and existing facility operations.

- Partner with universities and the international scientific community to maximize benefits.
- Ensure adequate support for priorities by utilizing management systems that reflect state-of-the-art business practices.
- Earn confidence by making realistic claims and delivering on what is promised.

Success Indicators

- Quality of science and innovativeness of research, as indicated by favorable outside peer review and expert advisory committee reports.
- Sustained achievement in advancing knowledge, as indicated by the impact of knowledge gained in other scientific and technological fields, and the number of publications, citations, and awards generated by DOE-supported research.
- Optimal operation of major experimental facilities, as indicated by operating efficiency and performance benchmarking.
- Development of new technologies that advance fundamental research capabilities and reduce costs, as indicated by new scientific and technology programs that emerge from the research.

Goal 3

Construct leading-edge experiments and user facilities on schedule, within budget, and in a safe and environmentally responsible manner.

Strategies

- Ensure that all facilities are "best-in-class" by using total quality management benchmark processes and state-of-the-art management information systems.
- Develop an oversight process that involves affected parties in a team approach to facility review.
- Employ risk-based cost benefit analysis to help set priorities and make decisions.
- Involve the international community to develop a global research facility network.
- Use innovative technologies to reduce costs.

Success Indicators

- Preeminence of facilities, as indicated by support by DOE and users, comparison with other facilities worldwide, the nature and extent of university and industrial involvement, and investment by users in the facility.
- Improved performance of facilities, as indicated by meeting original target performance plans and meeting expectations of users.
- Achieving construction cost and schedule milestones, as agreed upon prior to construction.
- Establishment and documentation of methods for determining and ensuring the level of and compliance with environmental, safety, and health standards.

Goal 4

Add value to the U.S. economy through the application of new and improved technologies.

Strategies

- Strengthen alignment between DOE programs and industry needs. Adequately plan and fund partnerships with industry.
- Streamline and improve the technology transfer process and learn how to work better with small businesses.

- Provide adequate funding for partnerships with industry.
- Provide adequate support for applied research and technology development.
- Forge links with other agencies and academia to leverage research benefits and avoid duplication.
- Strengthen links between laboratories.
- Establish national laboratories as regional centers to stimulate industrial competitiveness.

Success Indicators

- Increase in number and magnitude of cooperative activities with industry.
- Increase in technologies developed and deployed as a result of partnerships with industry.
- Increase in number of new projects at companies and problems solved or avoided that can be attributed to interactions with DOE programs.
- Increased leverage of government dollars with private sector funding.

Goal 5

Help provide a technically trained and diverse workforce for the Nation and enhance American scientific and technical literacy, especially in energy, the environment, and the impact of science on the economy.

Strategies

- Increase DOE participation in pre-college mathematics and science and continuing education programs.
- Increase DOE programs for teachers and students in the Department's laboratories.
- Support and encourage greater involvement by DOE science and technology staff in educational and community outreach programs.
- Expand opportunities in science at an early age for traditionally under-represented groups.
- Provide scientific and technical energy information through dissemination mechanisms responsive to customer needs, such as teacher networks, use of electronic networks, public television, and outreach vans and buses.

Success Indicators

- Improved scientific literacy of the American public and workers and increased participation of traditionally under-represented groups in technical education programs.
- Improved technical effectiveness of DOE and contractor employees, as indicated by work performance and community outreach.
- Increases in level of customer demand for departmental information resources and more positive feedback from information users.

National Security

For almost fifty years, our national security has relied on the deterrent provided by nuclear weapons. The diminishing strategic military threat, due to the end of the Cold War and break-up of the Soviet Union, has provided the opportunity to redirect priorities from weapons production activities to other critical missions. At the same time, the Nation continues to rely on its nuclear deterrent, including nuclear powered warships, to fulfill critical national security missions. Their continued safe and effective operations are essential to national security.

DRAFT

PROJECT MEMORANDUM

RAND

Assessment of Fundamental Science Programs in the Context of the Government Performance and Results Act (GPRA)

Susan E. Cozzens

PM–417-OSTP

April 1995

Prepared for the Office of Science and Technology Policy

Critical Technologies Institute

ASSESSMENT OF PROGRAMS IN THE CONTEXT OF THE GOVERNMENT PERFORMANCE AND RESULTS ACT (GPRA)

This paper discusses fundamental science program evaluation in light of GPRA's requirements. It provides a general overview of GPRA, examines the requirements it places on research agencies, and then looks at specific performance indicators that agencies might use in responding to the GPRA requirement for summary indicators.

OVERVIEW OF GPRA

In the summer of 1993, Congress passed the Government Performance and Results Act (GPRA). Its purpose is "to improve the efficiency and effectiveness of Federal programs by establishing a system to set goals for program performance and to measure results."[1] Because the Act shifts how federal agencies manage programs "from an input focus to an emphasis on performance and results,"[2] it supplements the call for an emphasis on results in the National Performance Review.[3] GPRA grew out of several related government management practices, including the trend toward use of program goal-setting and performance measurement in state and local governments, and at the national government level in several foreign countries.[4] The Chief Financial Officers Act of 1990 also acknowledged the need for more attention to performance measurement.[5] Senator William Roth, who first introduced a similar bill in 1990 and who now heads the Governmental Affairs Committee which oversees its implementation, calls GPRA "the single most important piece of the puzzle" in improving government performance.[6]

---

[1] Report of the Committee on Governmental Affairs, United States Senate, to accompany S. 20, Government Performance and Results Act of 1993, USGPO, June 16, 1993, 103rd Congress, 1st Session, Report 103-58, p. 2. [hereafter, "Senate report"]

[2] Leon E. Panetta, "Memorandum for the Heads of Executive Departments and Agencies," M-94-2, Executive Office of the President, Office of Management and Budget, October 8, 1993.

[3] Vice President Al Gore, "From Red Tape to Results: Creating a Government that Works Better and Costs Less," report of the National Performance Review, Washington, DC, September 7, 1993.

[4] Senate report, p. 9.

[5] Senate report, p. 6. However, as the legislative reports on GPRA point out, neither the CFOs Act itself, nor its legislative history, provides elaboration on the phrase "systematic measurement of performance," and the instructions about performance measures in the Act refer only to financial measures for agencies with substantial commercial functions. Thus, the Senate report on GPRA concludes that "... the annual financial statement will provide a limited basis for addressing program performance," and that "... the CFOs Act provides insufficient emphasis for extending performance measurement across the full range of agency program activities." (Senate report, p. 6)

[6] Senator William V. Roth, Jr., "Improving Government Performance," speech given at the Brookings Institution, March 22, 1995.

The GPRA legislation

GPRA lists five purposes, briefly stated as

- improve the confidence of the American people in their government, by holding Federal agencies accountable for achieving program results,
- initiate program performance reform,
- promote a new focus on results, service quality, and customer satisfaction,
- help Federal managers improve service delivery,
- improve Congressional decision making with better information on the effectiveness and efficiency of programs, and
- improve internal management of the Federal government.[7]

To achieve these goals, GPRA calls for a consultative, iterative process of strategic planning and assessment of progress. It requires agencies to

- develop strategic plans prior to FY98, consulting with Congress in the process;
- prepare annual plans setting performance goals beginning with FY99; and,
- report annually to OMB and Congress on actual performance compared to goals. The first report is due in March, 2000.

The law attempts to improve program management directly through the process of producing performance goals and measures, and to improve budget allocation by taking performance information into account. It does not set up performance budgeting across the government, although it does require pilot attempts in a few agencies to specify the levels of results expected at different budget increments.

Under GPRA, by September 30, 1997 (that is, with the FY99 budget submission), each federal agency is required to submit a strategic plan to OMB. The strategic plans are to include:

- a comprehensive mission statement;
- general goals and objectives for the agency's major functions;
- a summary of the resources, systems, and processes that are critical to achieving these goals;
- a description of how the general goals and objectives will be achieved; and
- a description of key external factors that could affect achievement of these general goals.

Also included is a description of how program evaluations are used in establishing goals, along with a schedule of future evaluations. The strategic plan is to cover at least five

[7] Government Performance and Results Act of 1993, 103rd Congress, 1st session, Report 103-106, Part I, p. 2. [hereafter, "GPRA"]

years forward from the fiscal year in which it is submitted, and is to be updated at least every three years.[8]

Beginning with FY99, the Act requires federal agencies to prepare annual performance plans for each program activity. A "program activity" is defined in the Act as "a specific activity or project" as listed in the Federal budget.[9] The annual performance plans are derived from the strategic plan and set specific performance goals for a fiscal year. The individual agency performance plans will be used to prepare a federal government performance plan. This overall plan is to be part of the annual Budget of the United States Government.[10]

In the annual performance plan, performance goals are generally to be expressed in an "objective, quantifiable, and measurable"[11] form, through performance indicators that measure or assess the relevant outputs, service levels, and outcomes of each program activity. The plan must also describe the means used to verify and validate the measured values.[12] If a performance goal cannot be expressed in an objective and quantifiable form, an alternative descriptive form may be used. But the indicators must provide a basis for comparing actual program results with the established performance goals.[13]

The Act establishes some common vocabulary for discussion of program performance. Implicitly, GPRA treats government activities and spending as *inputs* to a chain of activities that eventually produce benefits for the public. Government inputs are intended to produce both short-term *outputs* and longer-term *outcomes*.

- The Act defines an *output measure* as the tabulation, calculation, or recording of activity or effort.
- An *outcome measure,* as defined in GPRA, is an assessment of the results of a program activity compared to its intended purpose.

To use the example given in the legislative report on the Act, eligible clients completing a job training program are outputs; an increase in their rate of long-term employment is an outcome.

The guidance accompanying the Act explains that "output measures are often intermediate, in that they assess how well a program or operation is being carried out

---

[8] This paragraph is taken virtually verbatim from Panetta, October 8, 1993, op. cit.

[9] GPRA, op. cit. See next section for examples from research agencies. OMB has further identified a "program activity" as a major function or operation of an agency, or a major mission-type goal that cuts across agency components or organizations (Philip Lader, "Nomination of Pilot Projects," OMB memorandum, October 13, 1993)

[10] Panetta, October 8, 1993, op. cit.

[11] GPRA, p. 3.

[12] Senate report, p. 30. The use of audits is explicitly not required

[13] GPRA, p. 3

during a particular time period ... Output measures in performance plans should emphasize those used by agency officials in day-to-day operations and program management."[14] The report also acknowledges that "outcome measurement cannot be performed until a program or project reaches either a point of maturity (usually at least several years of full operation for programs continuing indefinitely) or at completion. Another prerequisite for measuring outcomes is the existence at the outset of a clear definition of what results are expected from the program or project. While recognizing that outcome measurement is often difficult, and is infeasible for some program activities, the Committee views outcome measures as the most important and desirable measures, because they gauge the ultimate success of government activities."[15]

Outputs and outcomes, then, are the short- and long-term indicators of program performance.

- A *performance goal*, as defined in the Act, is the target level of performance expressed as a tangible, measurable objective, against which actual achievement can be compared. For example, a performance goal for a student reading program is to have 2.3 million students receive an average of three additional hours of reading instruction per week during the 1990 school year.[16]
- A *performance indicator* is a particular value or characteristic used to measure output or outcome. In the previous example, the indicator is hours of reading instruction per week.

OMB has informed agencies that while the goals and indicators should be primarily those used by program managers to determine whether the program is achieving its intended objectives, they should also include measures that will be useful to agency heads and other stakeholders in framing an assessment of what the program or activity is accomplishing.[17]

GPRA stresses multiple performance indicators, and emphasizes outcome, rather than output, measures of performance. The report on the bill states "The Committee believes agencies should develop a range of related performance indicators, such as quantity, quality, timeliness, cost, and outcome. A range is important because most program activities require managers to balance their program's priorities among several

[14] GPRA, p. 3.

[15] House report, p. 19.

[16] Senate report, p. 32.

[17] Leon E. Panetta, "Memorandum for the Heads of Departments and Agencies Designated at Pilot Projects under P.L. 103-62," Executive Office of the President, Office of Management and Budget, March 3, 1994.

sub-goals.... While the Committee believes a range of measures is important for program management and should be included in agency performance plans, it also believes that measures of program outcomes, not outputs, are the key set of measures that should be reported to OMB and Congress."[18]

The challenge in responding to GPRA is that indicators are always partial, capturing some aspects and not others of the phenomenon of interest. Even a set of performance indicators provides only an approximate representation of a program's actual performance. Later parts of this paper discuss the correspondence between research performance indicators and research program performance.

The legislative report on GPRA considers the issue of the cost of implementing the Act. Most witnesses in hearings on the bill testified that the cost need not be burdensome, but the Act includes a requirement that costs be tracked and reported to Congress. An interagency group developing performance measures for research programs recommends that the data to be gathered should be valuable enough to program managers themselves that they are willing to spend what is needed. GPRA offers a reward in exchange for the effort invested in performance reporting. Eventually, in exchange for greater accountability, agencies will be allowed to waive administrative requirements and controls to provide greater managerial flexibility.

Relationship to program evaluation

GPRA distinguishes between the system of annual performance reporting it mandates and another, related activity, *program evaluation.*

- A *program evaluation* is an assessment, through objective measurement and systematic analysis, of the manner and extent to which Federal programs achieve intended objectives.[19]

The legislative report on the Act explains that "while most often aimed at assessing the degree to which a program's stated objectives are being or have been realized, program evaluations are also frequently used for measurement of 'unintended effects,' good or bad, that were not explicitly included in the original statement of objectives or foreseen in the implementation design. Thus, they can serve to validate or find error in the basic purposes and premises that underlay a program or policy."[20] GPRA assumes, rather than mandates, that an agency has an active effort underway in program evaluation.

---

[18] House report, p. 17.
[19] GPRA.
[20] House report, p. 20.

Federal program evaluation has a history of several decades.[21] In this tradition, program evaluation is a set of management practices that go far beyond performance plans and reporting. According to Elinor Chelimsky, " ... program evaluation is the application of systematic research methods to the assessment of program design, implementation, and effectiveness."[22] Six general categories of program evaluation have been distinguished:

- *Front-end analysis (preinstallation, context, feasibility analysis)*, which confirms, ascertains, or estimates needs, adequacy of conception, operational feasibility, etc. of the program.
- *Evaluability assessment,* including activities undertaken to assess whether other kinds of program evaluation efforts should be initiated.
- *Formative (developmental, process) evaluation*, which tests or appraises the process of an ongoing program in order to make modifications and improvements.
- *Impact (summative, outcome, effectiveness) evaluation,* which finds out whether a whole program works, and is intended to provide information useful in major decisions about program continuation, expansion, or reduction.
- *Program monitoring*, ranging from periodic checks of compliance with policy to relatively straightforward tracking of services delivered and counting of clients.
- *Evaluation of evaluation (secondary evaluation, metaevaluation, evaluation audit),* ranging from professional critique of evaluation reports to reanalyses of original data to summarize results across individual evaluations.[23]

In these terms, GPRA primarily calls for program monitoring and impact assessment--or, in the GPRA terminology we use in this paper, reports of output and outcome measures.

In GPRA, program evaluation plays a different role from performance plans and indicators. Program evaluations are more in-depth studies of program results, and are therefore usually done less frequently and more selectively than performance reporting.

---

[21]The published literature on general program evaluation is voluminous. One good recent overview is provided in Shadish, William R., Jr., Thomas D. Cook, and Laura C. Leviton, *Foundations of Program Evaluation: Theories of Practice,* Newbury Park, CA, Sage Publications, 1991. Cook and Shadish have also written an excellent short summary of the lessons learned from several decades of program evaluation: "Evaluation: The Worldly Science," *Annual Review of Psychology,* Vol. 37, pp. 193-232. Program evaluation in the government context described in Rist, Ray C., ed., *Program Evaluation and the Management of Government: Patterns and Prospects across Eight Nations,* New Brunswick, NJ, Transaction Publishers, 1990; and in Wye, Christopher G., *Evaluation in the Federal Government: Changes, Trends, and Opportunities*, San Francisco, Jossey-Bass, 1992.

[22]E. Chelimsky, "Old Patterns and New Directions in Program Evaluation," in *Program Evaluation: Patterns and Directions,* E. Chelimsky, ed., Washington, DC, American Society for Public Administration, 1985, p. 7.

[23]Taken almost verbatim from Peter H. Rossi (ed.), *Standards for Evaluation Practice*, San Francisco, Jossey-Bass Inc., 1982, pp. 9-10.

Program evaluation develops the knowledge base within an agency about outputs and outcomes of its programs, and thus helps research managers and agency executives assure that the agency's programs are effective mechanisms for reaching its goals. Program evaluation often develops output and outcome indicators, but interprets them in a descriptive framework. Agencies can draw GPRA summary indicators from among those developed in detailed program evaluation. But because GPRA performance indicators are aggregated across programs and need to be gathered and reported annually, as a practical matter they cannot reflect as much depth as the data and information used for a full-blown, detailed program evaluation.

OMB Guidance and the GPRA industry

GPRA mandates that the General Accounting Office monitor and report on the implementation of GPRA, and assigns responsibility for the implementation itself to the Office of Management and Budget. A small staff at OMB provides general guidance and facilitates discussions on the details of implementation, such as report formats. One of the first jobs of this staff was to designate GPRA pilot projects, as required in the Act. The Act specified that several sets of pilot projects were to be undertaken over the late 1990s, including a set to test and demonstrate to agencies how to produce annual performance plans and reports.[24] The Act calls for the pilot projects to be representative of major government functions and activities. The first set of designated pilot projects included programs from twenty agencies. Interaction with agencies on the pilot projects has been left in the hands of individual budget examiners, who will also be the first point of contact on full GPRA reporting requirements.

The first set of performance plans submitted under the GPRA pilot projects was an important learning experience for OMB.[25] Twenty percent of the plans were exemplars, demonstrating that measurable, quantitative performance goals could be set in advance. Another twenty percent, however, "lacked goals or measures with sufficient substance to be useful in managing a program, measuring performance, or in supporting a budget request. Put another way, if this were CY 1997 [when the whole government is required to submit plans], little or nothing worthwhile could be salvaged by OMB from agency plans such as these. ... A repeat of this experience three years hence would be a

---

[24] Research and development is among the 30 major programs, functions, and activities OMB identified for the purposes of judging whether its set of pilot projects was balanced (Lader, op. cit.). Three research-relevant pilot projects are discussed below. The summary and information on pilot projects are from Panetta, October 8, 1993, op. cit.

[25] Walter Groszyck, "Assessment of FY 1994 GPRA Pilot Project Performance Plans," OMB memorandum, August 10, 1994.

major blow to successful implementation of GPRA." The conclusion from this exercise was that "...the rest of the government needs to be engaged early-on if useful plans are to be forthcoming in 1997." OMB has therefore included GPRA-like activities in its spring review for the Fiscal Year 1997 budget, an exercise which is now beginning. It has asked agencies to designate "key programs," and will discuss performance measures and the availability of data on those programs with the agencies. The key programs will be expected to report performance measures with the FY97 budget request in September, 1995. In addition, with the FY97 budget, agencies are asked to name performance measures, although not to set performance goals, for all their programs.

As OMB's activity around GPRA has increased, attention to the Act has spread, and GPRA-related activities have sprung up within and outside government. A listing for February through July, 1995, for example, includes 47 activities, ranging in cost from $0 to $1795 per person.[26] The Office of Personnel Management hosts a GPRA Interest Group monthly, the Federal Quality Institute is offering a variety of classes, and a range of private groups are advertising specialized sessions on GPRA and related topics.

GPRA AND RESEARCH PROGRAMS

Program evaluation in research agencies

The evaluation of federal research programs has historically not been closely linked to the larger stream of government program evaluation, but has instead grown up as an independent tradition. Research program evaluation, like general program evaluation, is a learning process involving both program participants and other stakeholders in an in-depth look at how a program is working. It analyzes the objectives, priorities, and customers for the program; examines the structure of the program's portfolio; and considers the costs of the program in relation to its results. Good research program evaluation is done by independent evaluators, and includes assessors with relevant technical expertise and experiences in the type of research being evaluated as well as assessors from outside the research community. It gathers systematic evidence on program performance and relies on multiple lines of evidence to draw its conclusions, which are reported to program managers and participants, other stakeholders, and the public.[27]

---

[26] "Update on GPRA-related meetings and conferences, 4/7/95," Management Systems International, Inc., Washington, DC.

[27] This paragraph is drawn from: Practitioners' Working Group on Research Evaluation, "Evaluation of Fundamental Research Programs: A Review of the Issues," prepared at the request of the

U. S. research agencies have generally followed one of two approaches in their evaluations.[28] One is technical review by a panel of external experts, always including researchers and sometimes including users of research results as well. For example, since the 1950s, the intramural programs of the National Institute of Standards and Technology have been evaluated with extensive site visits by expert panels organized by a Board of Assessment which is a branch of the National Research Council. In the same spirit, the Office of Energy Research has a highly structured retrospective process of expert assessment at the project level, with panel scoring on pre-set criteria. The scores are aggregated at program level and reported within the agency.

A second approach to research program evaluation relies more extensively on data gathering by external contractors. Such evaluation studies, which draw more directly on the general program evaluation tradition, often use mail or telephone surveys or publication-based indicators, sometimes in combination with expert judgments of various sorts. An example is the National Science Foundation's mail survey of participants in its Research Experiences for Undergraduates program. Similarly, to assess prospects for collaboration with industry, the National Institute of Dental Research conducted an extensive study in the area of restorative dental materials research, using publication-based indicators, patent indicators, surveys, and case studies.

Evaluation studies, however, are relatively rare, and are concentrated in the fundamental science agencies, NSF and NIH. Most of the assessment of federal research programs is descriptive, and far removed from the sort of quantification of performance GPRA is seeking. A large array of quantitative tools for evaluation has been described in the literature,[29] but few of them are used in practice. To respond to GPRA, research programs and agencies thus face the challenge of either choosing among an array of options they have largely avoided in the past, or developing new ones.

GPRA reporting units

The option chosen for level of reporting will depend on agency goals and structures. In research, the range of activities covered under GPRA will be quite diverse and the level of aggregation for GPRA reporting will vary widely. As discussed earlier, GPRA requires agencies to set performance goals and report performance indicators for "specific activities or projects" as listed in the annual U.S. budget, but leaves the option to

---

Office of Science and Technology Policy, August 1, 1994. Available from the Critical Technologies Institute as part of the Metrics of Fundamental Science project.

[28] See Susan E. Cozzens, "U. S. Research Assessment at the Crossroads," Part I of a final report on NSF Grant SBE-9220059, August, 1994.

[29] We discuss these measures in the last section of this chapter.

"aggregate, disaggregate, or consolidate program activities, except that any aggregation or consolidation may not omit or minimize the significance of any program activity constituting a major function or operation for the agency."[30] For research programs, the choice of aggregation approach may be critical to the issue of developing realistic performance concepts.

The default option for units of aggregation is the budget line item; that is, agencies would prepare performance plans and report performance indicators for each item that appears as a line item in the government budget. Budget line items submerge research functions thoroughly into larger units in some agencies, but in others pit one program against another.[31] For example, at the Department of Defense, basic research, applied research, and development for the Department as a whole appear as line items in the Department's budget. Thus, under the default, Defense would need to present aggregate indicators for basic research in all three service branches together. Within each branch, research results would also be included in performance indicators for the service as a whole--that is, for example, as part of the Department of the Army for the Army Research Laboratory ($210 million), and as part of the Navy for the Office of Naval Research ($425 million). At the Department of Energy, performance indicators would be presented for the combined total of Atomic Energy Defense Activities, General Science Programs, and Energy Programs ($1.6 billion in basic research), but not separately for Basic Energy Sciences ($619 million). The Agricultural Research Service, which is a separate line item ($672 million, including $363 million in basic research) at Agriculture, would report its own indicators.

The three other large research-supporting agencies report more detail in their budgets. Therefore, if the default option is taken, at the Department of Health and Human Services, the National Institutes of Health would report in terms of its individual institutes (Cancer, Eye, Aging, etc.).[32] Likewise at NASA and NSF, programs in various areas of science would appear side-by-side in the performance indicators. For NASA, space science ($1.9 billion), life and microgravity science ($608 million), and Mission to Planet Earth ($1.3 billion) would each report indicators, as would NASA's academic programs ($104 million). At NSF, the various directorates--geological sciences and biological sciences, for example ($434 million and $314 million respectively)--would report indicators, as would the Social and Behavioral Sciences Directorate ($96 million),

[30] GPRA.

[31] These numbers are based on FY95 budget authority and the Budget of the United States for that year.

[32] This discussion is based on data from the FY95 budget, and does not reflect any reorganization of HHS after the departure of the Social Security Administration.

which includes the division that handles international programs and the one that produces science statistics.

Alternatives can be proposed, however. The National Science Foundation is considering reporting in "portfolios" of programs and activities that have similar performance concepts, for example, NSF's entire portfolios of centers, facilities, or project grants.[33] To make a statement about how effective NSF as a whole is, it is easier to summarize sets of unique indicators for such areas than to sum up across discipline-based directorates. As mentioned earlier, OMB has also opened the door for agencies to report on "missions" that cut across their functional units. So, for example, if it chose to do so, the National Institutes of Health could report on its entire cancer effort, intramural and extramural, located in and funded out of many different institutes, ranging from fundamental molecular biology to clinical trials. The only level of organization seemingly ruled out for GPRA purposes is cross-agency programs, since performance reporting is at the agency level. But agencies may elect to report on their contributions to interagency initiatives, and such reporting may be required for other purposes, such as OSTP's response to requests from the Senate Appropriations Subcommittee (discussed in another paper prepared for this project).

This discussion of levels of aggregation for performance indicators illustrates again the differences between regular program evaluation and performance reporting under GPRA. The default units of aggregation are much too broad to be the object of program evaluation, which would focus instead on more coherent sets of activities further down in the organization. Some of the alternative units suggested here might be the focus of a special evaluation, but this would certainly not happen on an annual basis. Thus we see again how program evaluation and GPRA-type performance indicators complement each other in the overall research management process.

Research performance concepts

The fundamental difficulty for research programs in responding to the performance reporting requirements of GPRA, especially at such broad, summary levels, is the character of research goals. The outputs of research programs are tangible and measurable, but the outcomes are less so. Furthermore, the connection between the two varies widely, even among federal programs of fundamental research. Therefore, in performance indicators, one size is unlikely to fit all.

For all science programs, the primary goal is to produce knowledge that will be used in service to society. But the programs vary a great deal in how close or far away

[33] See discussion below of NSF's pilot project in facilities indicators.

their research activities are from practical application. At one end of the spectrum, the National Science Foundation's primary mission is to "enable the United States to uphold a position of world leadership in all aspects of science, mathematics, and engineering."[34] NSF's research, therefore, must first and foremost be of world-class caliber, and also balanced among fields. At another point on the spectrum, the Agricultural Research Service mission is to "develop new knowledge and technology needed to solve technical agricultural problems of broad scope and high national priority... "[35] ARS's research must therefore first and foremost contribute to solving agricultural problems. In both cases, the knowledge outputs of the programs take the same form: research results, communicated to various audiences through workshops, conferences, reports, and publications. The outcomes to which the research results contribute, however, are quite different--world leadership versus agricultural problem solving--and call for different kinds of performance indicators.

At the same time, research programs build technical capacity by investing in human resources. Within a mission-oriented government laboratory like the Army Research Laboratory or the National Institutes of Standards and Technology, human resources are appropriately treated as an input to the research process. But agencies that primarily support extramural research develop human capital as a generic national resource: trained people are an output in these cases. NIH and NSF are prime examples. By supporting research at universities, these agencies invest in two sets of people: the principal investigators themselves, who are kept at the frontiers of knowledge through research activity; and also new Ph.D.s and the other professionals trained in part by the PIs, for example, medical students taught by NIH-supported investigators, or undergraduate engineers taught by NSF-supported engineering PIs. The expertise embodied in these people is employed in service to society far away from the funding organization, in transactions that are not necessarily connected to the grant the organization provided. So for example, an ecologist supported by NSF early in her career may eventually head a branch of the Forest Service, or a neuroscientist supported by the National Institute on Aging may contribute to drug development by consulting with a pharmaceutical firm years later. While trained people are visible outputs of the research projects the agencies support, the longer-term outcomes of those investments are seldom visible, especially at the end of the project period.

To summarize, outcomes produced by both increases in understanding and investments in human capital are frequently long-term, and they are mediated through

[34] National Science Foundation Strategic Plan, February 1995.
[35] Agricultural Research Service Program Plan, October 1991.

institutions which use the knowledge base and trained people the program produces. Outcomes could be tracked over a long period of time starting from individual programs, but by the time many of them appeared, the direct effects of the program would have interacted with many other factors. The problems of connecting research to outcomes are shared across fundamental research programs.[36]

The GPRA requirement for regular program evaluation at a more detailed level only partially alleviates this problem. Program evaluation can go into more depth, use more indicators, and construct a more complex picture of the program than an annual summary performance report can. It can draw on knowledge of long-term processes produced from retrospective studies, like the ones described in Section IV of this report. Program evaluation also inevitably involves interpretation of indicators by people who are knowledgeable about the program and how it works: program managers, evaluation staff in agencies, steering committees organized specifically for the evaluation, and agency decision makers. In this context, the use of quantitative indicators appears both rational and wise, since they can supplement descriptive expert judgment without supplanting it.

In the context of an annual summary performance report, however, the descriptive elements may fall away. The readers of the report are likely to know much less about the program, and will not be able to supply the descriptive information themselves. For summary performance reports, then, indicators have to be chosen with special care, and packaged as much as possible with the other kinds of information necessary to understand the program. The interagency group discussing research performance measures has therefore stressed that to minimize the likelihood of misinterpretation, the indicators should always be reported in narrative form, and examples of research advances will generally be necessary to supplement quantitative indicators.[37]

An example of how to package such information comes from the other side of the Atlantic. To convey performance information to its relevant ministry, one Swedish research council has developed a one-page format that meets the description offered by the U.S. interagency group. The Swedish Board for Technical Development (NUTEK), responding under a law very much like GPRA, has chosen to prepare one-page, two-column abstracts for each of its programs, with text that describes the program goals (in 4-5 lines of description), the actual work done under the program, the number of projects

[36] W. Herman and M. Bosin of the Food and Drug Administration have distinguished usefully between quantitative input-to-outcome models and nonquantitative (logic-flow) models in the analysis of outcomes from research programs. They have presented this work to the Roundtable on Research Performance Measures, and have viewgraphs available.

[37] See minutes of Roundtable on Research Performance Measures.

supported, steering committee members, outputs (including doctorates awarded, publications, start-up companies, new lines of research in existing companies, etc.), and NUTEK's subjective views on future outcomes of the program. The experience thus far is that such reports are enthusiastically received at the Ministry, where a brief, readable format for solid information about the program is useful for several purposes.[38] The reports also increase the likelihood that quantitative indicators will be used in context rather than alone.

GPRA R&D PILOT PROJECTS

The R&D pilot projects under GPRA are already struggling with the tensions inherent in the tasks the Act requires. While it is not yet clear whether all the efforts under the pilot projects will translate into usable summary performance indicators at agency level, much has been learned in the pilot project process.

The National Science Foundation houses four pilot projects, but one focuses on an administrative activity and will not concern us here. The other three represent three different kinds of NSF programmatic activity:

- the Science and Technology Centers Program, representing a funding mode that contributes to three different NSF strategic goals
- a set of user facilities (telescopes, synchrotrons, etc.) in the Mathematics and Physical Science Directorate, and
- the High Performance Computing and Communications strategic area.

The Army Research Laboratory, which supports fundamental and applied research, was included as a second round pilot.

NSF's Science and Technology Centers

NSF's Science and Technology Centers (STC) program activity provides stable, long-term support to 25 university-based research centers.[39] As an experimental mode for federal support of scientific research, the activity seeks to establish the potential benefits of stable long-term support for interdisciplinary research and activities in a center setting. Its goals are

- to address interdisciplinary research problems beyond the capabilities of single investigators or small research groups;

[38]Cozzens interview with Torbjorn Winqvist, NUTEK, September 1994.

[39]Information here provided to this project by David Schindel, STC Program Manager, September 1994, and drawn from the STC pilot project proposal.

- to promote partnerships with states, industry, and national research laboratories, for knowledge and technology transfer from academia to other sectors; and
- to produce graduates at all levels with unique interdisciplinary science and engineering training.

As a GPRA pilot project, the STC program had the advantage of drawing on a workshop held in January, 1992, at which center directors and professional evaluators explored the topic of appropriate performance indicators for center programs. To develop its first pilot project performance plan, then, the STC drew on the performance concepts carefully developed in this context, and convened an in-house advisory group of program evaluation specialists who contributed ideas and helped to refine the pilot project's FY 1994 performance plan. The performance indicators that were proposed are shown in Table 2.1. The program proposed to verify and validate these indicators through the use of an independent evaluator and expert panels.

In its response to this first performance plan, OMB pressed the program to choose indicators that were more quantitative in nature. OMB also raised the question of how a baseline for measuring future performance can be established with indicators of this sort, and wanted to know how to judge a good versus a poor level of performance on these indicators. In preparing the pilot project's FY 1995 performance plan, the program then decided to follow an alternative path offered in GPRA. GPRA states "if an agency, in consultation with the Director of the Office of Management and Budget, determines that it is not feasible to express the performance goals for a particular program activity in an objective, quantifiable, and measurable form, the Director of the Office of Management and Budget may authorize an alternative form. Such an alternative form shall--

(1) include separate descriptive statements of--

(A) (i) a minimally effective program, and

(ii) a successful program, or

(B) such alternative as authorized by the Director of the Office of Management and Budget,

with sufficient precision and in such terms that would allow for an accurate, independent determination of whether the program activity's performance meets the criteria of the description; or

(2) state why it is infeasible or impractical to express a performance goal in any form for the program activity[40]

The FY 1995 Performance Plan that was proposed by the STC pilot project and adopted is in fact very complex, but follows two simple principles.

[40] Government and Performance and Results Act of 1993, Public Law 103-62, Section 4.

- *First, the plan describes "significant progress" and "outstanding progress" with regard to two indicators in each of three program goal areas (research, education, and knowledge transfer). The Center is considered to have reached a particular goal if it is making outstanding progress on one of the two performance indicators, or significant progress on both indicators.* The two levels are in general distinguished by different adjectives, for example, one of the performance indicators for the goal of knowledge transfer focuses on collaborative partnerships. Significant progress using this indicator is defined, in part, as having a center "open to and used by at least a modest number" of researchers from other institutions, whereas in "outstanding progress" requires, in part, that such collaborators are "routinely integrated" into the center and "constitute a visible and regular component of the Center's population." For its central research indicator, "significant progress" is defined in terms that are not readily quantifiable: a majority of the Center's research products "appear in the field's most respected peer-reviewed vehicles." But "outstanding progress" is achieved through breakthroughs, when several of the Center's research products "are counted among the most influential contributions affecting the current direction of the field." For these and other indicators, the judgment of specialists from outside the Centers themselves is utilized.
- *Second, the plan treats the set of centers supported under the program as a portfolio, and sets performance goals for the program in the form of the percent of centers that match certain performance characteristics.*

For example, the program would be considered minimally if half its constituent centers are successful in reaching two or more goals or 80% are successful in reaching one goal. The program would be fully successful if 90% of the centers reach one goal, 75% reach two, or 20% reach three. Treating the center program as a portfolio encourages risk-taking and creativity. The program can succeed even if individual centers fail..

Table 2.1

Performance Indicators Proposed

for the Science and Technology Centers Program, NSF

Goal 1: *Address interdisciplinary research problems beyond the capabilities of single investigators or small research groups.* Indicators:

- amount, quality, and interdisciplinary nature of research products, such as publications and collaborations;
- novel integrative approaches to research problems;
- use and development of unique research instruments and facilities;
- size and duration of collaborative research projects; and
- sponsorship of leadership and coordinating activities for the community.

Goal 2: *Promote partnerships with states, industry, and national research laboratories, for knowledge and technology transfer from academia to other sectors.* Indicators:

- creation of mechanisms for knowledge/technology transfer: alliances, contractual arrangements, organizational structures
- tangible products of knowledge/technology transfer such as
    - amount, nature, and quality of joint research and personnel exchange;
    - patent disclosures, licensing arrangements;
    - new products, spin-off companies;
    - use and acquisition of shared research instrumentation; and
    - reciprocal training programs and hiring practices.

Goal 3: *Produce graduates at all levels with unique interdisciplinary science and engineering training.* Indicators:

- number of students trained, including minorities and females;
- inter-departmental/interdisciplinary nature of advising committees;
- career tracks of graduates;
- creation of interdisciplinary courses, curricula, majors, programs.

NSF facilities pilot

NSF is also using the portfolio concept in its pilot project on facilities, although in a different format than in the STC pilot. The Foundation supports a number of user facilities, such as telescopes and accelerators, in several different disciplinary directorates. The facilities have a common purpose: as phrased in the NSF strategic plan, "to enable the United States to uphold a position of world leadership" in selected fields of science. Each facility is planned in response to needs in a specific field, and each one starts from a different technical baseline. In the first performance plan, the participants in the pilot project developed five generic goals for their facilities, but had

difficulty translating them into practical performance measures and indicators. When facilities directors were asked to produce whatever data they thought was useful in relation to these five goals for the project's first performance report in relation the first plan, some interesting ideas emerged. The pilot project leader then called the group together, along with some representatives of other major facilities NSF supports, and the group developed a dozen generic performance indicators under three broad performance concepts, as indicated in Table 2.2.

The key to the plan was to think about the performance measures in terms of percentage change from a baseline. The baseline number could be different for each facility, and even measured in different metrics. To help standardize, the group had to invent a term, "user units," to refer generically to entities like beamtime and observing hours. In the end, however, the percentage change from each facility could be folded into a percent change for the whole portfolio of facilities. As with the STC pilot, the portfolio concept allows for variation among the facilities in their indicators for any particular year. When an individual facility experiences bad weather, for example, its figures may drop; but that individual variation will play only a small role in the Foundation-wide average.

Thinking in terms of the portfolio concept was a major challenge for the group because NSF normally devotes so much attention to the evaluation of individual facilities. The performance indicators for the individual facilities would of course be available to program managers and site visit teams, but when used at that level they would be interpreted in context, with regard to the performance expectations for that particular site. At the same time, improving the generic, aggregate performance characteristics of the portfolio can become a goal for the Foundation as a whole.

Table 2.2

Illustrative Performance Goals
from the NSF Facilities Pilot

Goal A: Efficiency of Operations

Facilities operate efficiently by ensuring that construction, upgrades, maintenance, acquisition and development of major instrumentation remain on schedule and within budget...

- Fraction of total originally scheduled user units [e.g., beam time, observing hours] lost due to breakdowns or other circumstances considered within the

control of the facility. *The objective is that this fraction not exceed 5% of the total originally scheduled user units.*

- Percentage change in the cost per available user unit from the previous year (calculated in constant dollars). *The objective is that this percentage change be at least -1%.*

Goal B: Effectiveness of Operations -- Scientific User Community

Facilities serve a broad user community effectively by providing for the development of major research instrumentation and facility upgrades that broaden research opportunities and encourage technology breakthroughs. The facilities remain at the cutting edge through their research and instrumentation development programs.

- The demand on the facility by users as indicated by the over subscription rate on the facility services and the percentage change from the previous year. This rate will be calculated as the fraction of user units requested compared to the total number of user units available. *Our initial objective is to see the over subscription rate decrease through increases in capacity or in efficiencies of equipment.*- The fraction of user units used by industry compared to the total number of user units available and the percentage change from the previous year. *Our initial objective is for this percentage to increase.*

Goal C: Effectiveness of Activities -- External Community

Research carried out in whole or in part through the facilities is demonstrably world class, with broad promotion of the potential uses of the research results and any technology breakthroughs. The facilities are substantively involved in the education and training of science and engineering students, and enhance the public awareness of science and the goals of scientific research.

- The fractional number of undergraduate and graduate students using the facility compared to the total number of users conducting research on the facility and the percentage change from the previous year. *Our initial objective is for this percentage to increase.*
- Number of accesses to the facility's World Wide Web service, and percentage change from the previous year. *Our initial objective is for this percentage to increase.*

The third NSF pilot project illustrates the difference between GPRA-type performance indicators and the "milestones" commonly requested under inter-agency programs. HPCC has been designated by the National Science Foundation as "strategic research in an area of importance to the nation." HPCC simultaneously combines activities in different NSF directorates, and contributes to an inter-agency initiative of the same name. The pilot project includes activities of several different kinds:

- fundamental research and cutting-edge technological development, both of which build a world-class knowledge base about computing and communications, and their utilization by the research and education communities;
- networking activities that focus on the very high bandwidth requirements of the research and education communities; and
- educational activities, focusing on special training at multiple levels at the cutting edge of computing knowledge.

At the core of HPCC activities is the effort to increase the nation's storehouse of computing science and information processing knowledge, and human capital. An important, but secondary aspect is the advancement of the infrastructure that enables the goal.

Some HPCC goals can therefore be phrased explicitly in terms of increased technical capacity. For example, in its FY95 performance plan, the program expected to "increase bandwidth of vBNS to about 600 Mb/second within two years," starting from a baseline of 155 Mb/second. Such goals are classical "milestones": they imply that the program has set a direction and provide concrete measures that can indicate whether it has moved toward it.

The development of more powerful computers is another example where technical capability can be increased and measured. The original Federal HPCC program for example has as a goal the demonstration of a sustained teraflop on Grand Challenge problems. However, a recent report from a National Academy of Science study on HPCC, recommended that teraflop computing should be treated as a direction, not a destination.

Thus, performance indicators for the program cannot focus only on increasing technical capacity, lest they conceal the more important aspects of the program such as increasing the pool of computing science and information processing knowledge, and human capital. The HPCC pilot project therefore includes a range of indicators, intended to capture both the fundamental increase in knowledge and the increasing technical capacity, and calls attention to the importance of unexpected results.

The Army Research Laboratory pilot

Like the NSF Science and Technology Centers, the Army Research Laboratory (ARL) had a head start on the problems posed by GPRA, in the sense that it had recently gone through a serious business planning process and had a strategic plan in hand with which to set performance goals.[41] ARL was eager to participate as a pilot because of the flexibility incentives GPRA offers in exchange for increased accountability. ARL thought that flexibility, particularly on personnel matters, would help it cope with the major downsizing it is experiencing, anticipated to be close to 40 percent in personnel over the 1990s. Over the last 20 years, study after study of Defense Department laboratories has called for greater flexibility and authority for laboratory management, to deal with a bureaucratic personnel and procurement system. ARL hoped that GPRA might be the mechanism for responding to these 20 years of concerns.

ARL rated the conventional tools of research evaluation as follows (see Figure 2.1). Long time-frame retrospective studies (like Project Hindsight, which tracked the origins of a set of Defense technologies in the 1970s), provide the most meaningful, but least timely information. Retrospective peer review provides meaningful, timely information that is limited in scope. The usefulness of metrics of inputs, surrogates, and outputs is timely, but varies in meaningfulness. Thus retrospective peer review, metrics, and customer satisfaction surveys all were seen to have different strengths and weaknesses.

ARL therefore decided to combine the methods in their plan. Their indicators list, shown in Figure 2.2, included 27 non-personnel indicators and 13 personnel indicators in early 1994. As with NSF, after seeing the first performance plan, OMB asked ARL to reduce the number of indicators. ARL first experimented with combining their list into aggregate measurements by asking higher level Army managers to assign weights, which were to be fed into a combination algorithm.[42] They are now moving toward more emphasis on retrospective peer review of a descriptive kind.[43]

Figure 2.1--The ARL Performance Evaluation Construct

Figure 2.2--Laboratory Performance Metrics

[41]Information here provided by Edward A. Brown, ARL, in a briefing for the Federal Research Assessment Network in May, 1994.

[42]T.L. Saaty, "*The Analytic Hierarchy Process*, McGraw-Hill International, 1980.

[43] Susan: check Roundtable notes on this.

THE CHALLENGE OF SUMMARY PERFORMANCE INDICATORS

If agencies choose the summary indicators option in GPRA (rather than the alternative format), then responding to GPRA will require selecting a judicious combination of indicators based on short-term program outputs that are significant because of their connection to longer-term processes. For constructing summary indicators, all available evaluation methods have both strengths and limitations. Many of the indicators of research program outputs could find useful applications in the context of a full-blown program evaluation, but have more severe limitations for use as GPRA summary performance indicators. For example, a full-blown evaluation can take into account descriptive analysis of interview data, complex models of program operation, or sophisticated citation analyses. All of these can provide performance-related information to inform an evaluation report, but do not match GPRA's requirements for simple performance indictors.

By definition, the primary goal of any research program is to increase understanding of a physical, social, or technological phenomenon. While understanding itself is hard to quantify, knowledge production has proven to be at least in part measurable. Three aspects of the knowledge produced under research programs are generally of interest to agency program managers: quantity, quality, and importance. This section reviews indicators of these three concepts, and mentions indicators of other aspects of the performance of research programs.

Quantity of Knowledge

Publications. Publication counts are by far the most widely used metric of knowledge production in science, finding applications from individual evaluation for promotion and tenure at universities to national science indicators.[44] European evaluations of university units have routinely included publication counts as one type of productivity index for a decade or so. In the British system, researchers asked for these more objective indicators to be included in the evaluation system to counteract arbitrary

[44]For detailed explanations of bibliometrics, see: A. F. J. Van Raan, 1993. "Advanced Bibliometric Methods to Assess Research Performance and Scientific Development: Basic Principles and Recent Practical Applications," review report based on invited paper, University of Leiden Report CWTS-93-05, August; Susan E. Cozzens, 1989. "Literature-Based Data in Research Evaluation: A Manager's Guide to Bibliometrics," report to the National Science Foundation; Francis Narin, Dominic Olivastro, and Kimberly Stevens, 1994. "Bibliometrics/Theory, Practice and Problems." *Evaluation Review*, Vol. 18, No. 1, February, pp. 65-76; A. F. J. Van Raan, 1993. "Advanced Bibliometric Methods to Assess Research Performance and Scientific Development: Basic Principles and Recent Practical Applications," review report based on invited paper, University of Leiden Report CWTS-93-05, August. For individual bibliometric methods, see notes on following pages.

judgments by parochial peer reviewers in a first round of university evaluations.[45] In the United States, publication counts were among the first evaluative indicators assembled at the National Institutes of Health and National Science Foundation.[46] Technical evaluation panels are often given publication lists for the researchers they are evaluating, for example, at the Office of Naval Research and in the evaluation process for intramural research at NIH.[47] Even programs that carry out no other evaluation activities often include number of publications in their lists of achievements.[48]

The use of publications as a metric of knowledge output has a long and respected history. It rests on a sociological theory that maintains that the norms of science require researchers to share their results with others in order to get credit for them.[49] However, differences in incentive and reward systems among the sciences and among research settings call for modifications in this theory. For university researchers, the norms of publication are undoubtedly strong. For those in other settings, publication may not be encouraged or may even be actively discouraged.

Some disciplines are also more publication oriented than others. Computer scientists, for example, often claim that programs are their major output, rather than publications. Publication counts are also limited in their applicability to cross-field comparisons because the "least publishable unit" varies among fields of science. Earth scientists publish their work in large chunks, incorporating great swaths of data in models and theoretical arguments. Laboratory scientists, from engineers through molecular biologists, may carve out smaller slices from their flow of work to publish. Social

---

[45]A. J. Phillimore, "University research performance indicators in practice: The University Grants Committee's Evaluation of British universities, 1985-86," *Research Policy* 18 (1989): 255-271; Ben R. Martin and Jim E. F. Skea, "Academic Research Performance Indicators: An assessment of the possibilities," Science Policy Research Unit, UK, March 1992. The UK funding councils recently decided to limit publication lists submitted by universities in resource allocation processes to the four best papers individuals in departments have published over the last three years. [See Claire O'Brien, "Quantity No Longer Counts in Britain," *Science* 264 (24 June): 1840, 1994.] It is now not the publication counts that figure in the resource allocation decisions, but rather the quality of the best publications, as judged by peers.

[46]NSF sponsored the first handbook in this area: F. Narin, *Evaluative Bibliometrics*, report on NSF contract C-637, March 31, 1976. NIH built an extensive bibliometric data base related to its programs in the 1970s and 1980s, reported in a series of institute-by-institute program evaluation reports.

[47]See R. N. Kostoff, "Evaluation of proposed and existing accelerated research programs by the Office of Naval Research," *IEEE Transactions in Engineering Management* 35:4, Nov. 1988; Report of the External Advisory Committee of the Director's Advisory Committee, National Institutes of Health, "The Intramural Research Program," Final Draft, April 11, 1994.

[48]For example, NASA's Microgravity Research Program.

[49]This is a brief statement of the theory of publication and reward put forward by Robert K. Merton and elaborated by his students.

scientists may wait and publish a book.[50] Collaboration patterns also affect the number of publications that appear in a field. Finally, the use of publication counts as performance indicators may skew the numbers upward, as researchers respond to this reward system.

As an output measure for research programs, then, publication counts may be a reasonable choice if the publication habits of the scientists supported by the program are fairly similar to each other, and if there is a stable core of researchers who work in settings that encourage publication. Programs that choose to use publication counts as indicators often place boundaries around the set of publications they will choose to count. They may, for example, limit the data to papers that appear in peer-reviewed journals, asking investigators to provide this information. Or they may choose only high-impact journals in the field, or only journals indexed in a prominent indexing service with good coverage.[51] Steps like these assure some homogeneity in the units being included in the metric.

Even after these caveats and corrective measures have been taken into account, however, there are inherent limitations in how much publication counts can say about the knowledge outputs of research programs. Fundamentally, publication counts are an activity measure, and leave out other important characteristics of the growth of knowledge among researchers supported by a program.

Other Output Measures. Other less widely applicable measures of activity or output are also sometimes used with regard to research programs, when they are deemed appropriate by program managers and participants. These include patents, devices, computer programs, and other signs of invention. ARL's list of indicators, for example, includes patents and invention disclosures. For research programs, such data are generally treated as a supplement to knowledge output indicators, but not the major indicator. When patenting activity resulting from basic research programs has been examined, the level of activity is often low.[52] Since small numbers are relatively unstable, including such a count in an aggregated set of performance indicators for a research agency is a risky strategy. In the context of detailed program evaluation, however, where a richer set of indicators is examined by a more knowledgeable group of

---

[50]John Ziman, 1994. *Prometheus Bound: Science in a Dynamic Steady State.* (Cambridge: Cambridge University Press); p. 104.

[51]The Engineering Research Council (TFR) in Sweden follows this strategy. One Dutch study also focused on the publications that appeared in top-ranked journals, rather than total publications, as an indicator of quality (Henk Rigter, "Evaluation of Performance of Health Research in the Netherlands," *Research Policy*, Vol. 15, 1986, pp. 33-48.

[52]For an example, see Research Corporation, "Study of Patents Resulting from NSF Chemistry Program," final report on NSF Contract EVL-8107270, New York, 1982.

people, even small levels of patent activity may be a relevant sign of certain kinds of important connections between research and the marketplace.

Quality of Knowledge

Researchers are usually more concerned with the scientific quality of the knowledge produced under a program than with its sheer quantity. Two major approaches to measuring quality appear in the literature: technical review and citation counts. Each is discussed below. Fortunately, in a large number of studies, their results have been found to be correlated for aggregates of publications (see discussion below). Awards and honorific positions have also been used as indicators of the quality of researchers supported by a program, and thus indirectly as an indicator of the quality of the knowledge they produce.[53]

Technical Review. Expert review is the most widely used approach in research evaluation, both in the United States and around the world. "The Nordic Model" of research evaluation, pioneered in Sweden and also used frequently in the other Scandinavian countries, uses small panels of international reviewers, who judge the national effort in a narrow field of science based on a week of site visits to the major laboratories.[54] Research managers have valued these visiting panels more for the place they fill in an overall research management system than for their specific results. For example, in a small country, a peer review panel that judges proposals can become ingrown, or be too soft on researchers who are no longer productive. An external panel that looks at the quality of the projects supported, or even the knowledge that an external panel will at some point be convened, can keep national review panels on their toes and strengthen their resolve with regard to weak research teams.[55] The Swedish research councils, however, have recently decided that they have learned about as much as they can from that system for the time being, and are experimenting with more comprehensive reviews of larger fields.[56] For over a decade, the European Community has also relied heavily on technical review to evaluate its programs. But as experience has accumulated,

---

[53]For an example, see "Sources of Financial Support for Research Prize Winners," National Science Foundation, NSF 87-87. ARL's indicators list includes significant awards, invited presentations, and prestigious posts such as journal editorships and officer ships in professional societies.

[54]E. Ormala, "Nordic experiences of the evaluation of technical research and development," *Research Policy* 18: 313-42, 1989; E. Ormala, "Impact Assessment: European Experience of Qualitative Methods and Practices."

[55]T. Luukkonen and B. Stahle, "Evaluation of Research Fields: Scientists' views," *Nord* 1993: 15, Nordic Council of Ministers, Copenhagen, 1993; Cozzens interviews with Swedish Research Councils, September 1994.

[56]Interview with Annette Wiklund, Swedish Natural Sciences Research Council, September 1994.

the system has come under criticism for relying too heavily on basic researchers to evaluate applied research programs, and the Community is now considering changes.[57]

In the United States, technical review varies among agencies from very informal assessment processes to highly structured retrospective quality control mechanisms. For example, at the informal end of the spectrum, the Agricultural Research Service examines the results of various aspects of its programs with workshops in various laboratories, attended by outside scientists and some users of research results. At the formal end (as mentioned earlier), the Office of Energy Research at the Department of Energy runs highly structured peer assessments of selected programs, evaluating hundreds of projects each year. In these assessments, the format is pre-established, and the reviewers rate the projects on standard categories. (See Figure 2.3) Within one review, then, the process transforms the descriptive judgments of peers into quantitative ratings, which can be compared across projects to identify those that need improvement.

There are known difficulties in structuring and using technical assessments, even for internal program evaluation purposes. There is no objective entity "quality" which can be measured objectively, as GPRA requires. Quality is a collective perception, and peer review panels have certain well-known limitations as representations of collective perception. In particular, the results of the evaluation are highly dependent on the choice of reviewers,[58] and cognitive particularism has been demonstrated--that is, biases of reviewers toward work of the type they are doing.[59] The practice of organizing a good technical review is designed to counteract these problems. Discussions of the state of the art in picking reviewers tend to stress first, getting a breadth of competence that matches the program well, and second, getting well-respected people so that the credibility of the review is established beyond doubt. Independence of reviewers is also considered essential for this purpose. Even after care is taken with these matters, however, technical review remains essentially a process of judgment.

Technical judgment processes would encounter additional difficulties if they were used to produce summary performance indicators to respond to GPRA requirements for annual summary indicators. One is their cost and intrusiveness. Current best practice for technical review involves face-to-face interaction between researchers and reviewers at a

[57]Hans Skoie, "EC Research and Technology Policies: Some Characteristics of Its Development and Future Perspectives," Institute for Studies in Research and Higher Education, Oslo, Norway, 1993; plus a recent communication with the Commission of the European Communities in Washington.

[58]Stephen Cole, Jonathan R. Cole, and Gary A. Simon, "Chance and Consensus in Peer Review," *Science*, Vol. 214, 1981, pp. 881-886.

[59]G.D.L. Travis and H.M. Collins, "New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System," *Science, Technology, and Human Values*, Vol. 16, No. 3, 1991, pp. 322-341.

fairly detailed technical level. If this method were applied annually to all, or even a sample of, federal research programs, the price in reviewer time alone would be enormous,[60] and would surely violate the principle of keeping GPRA implementation costs to a minimum.

In some cases, ratings might be gathered at no additional cost from expert panels already doing retrospective evaluation of projects or programs. Such retrospective peer review is quite common with regard to federal intramural laboratories and facilities supported extramurally. For example, NSF could ask the panels that evaluate its facilities for renewal funding to fill out forms rating the facilities on various performance characteristics and giving a summary rating. These ratings could be aggregated into a portfolio measure and added to the other data that the facilities have suggested reporting (see discussion of NSF facilities pilot project above). Then, instead of simply telling OMB and Congress that its facilities are evaluated by such teams, NSF could report the aggregate rating of the facilities examined in any particular year on a scale, perhaps from "world class" to "of marginal use." Such a rating would probably not convey new information to the facility or its program manager, but it might communicate the value of the facility better to outside audiences. At the very least, its marginal costs would be low.

Using peer ratings to perform comparisons across fields, however, raises as many methodological problems as the equivalent use of publication counts. Given the sensitivity of technical evaluations to the particular set of individuals involved on the review panel, the reliability of ratings from one year to the next in an annual process would be open to question under any circumstances. In the context of the budget process, however, where the GPRA performance indicators will be reported, the quantitative ratings provided by technical panels may be even less reliable. When technical reviewers are asked to produce ratings that lead to more or less money for their fields, they tend to skew their ratings upward. NIH has experienced this sort of rating inflation with regard to peer review ratings of proposals, and in fact experimented for a time with normalization systems to correct for such biases. Because of this potential problem, it would be risky at best to set baselines (as GPRA requires) or do comparisons between broad scientific programs (for example, among the three science areas NASA is likely to report on) based on technical review ratings alone.

The qualifications about technical review raised in this discussion relate only to constructing quantitative summary performance indicators for GPRA. The caveats here need to be carefully distinguished from the pros and cons of using descriptive technical

[60]Kostoff gives a generous estimate of the full costs of a typical technical review in his "Research Impact Assessment: Where are we now?" Summer 1994.

judgments either as part of the summary performance report for agencies, or in the full detailed program evaluation process. In both those other applications, expert judgments are considered essential.

Citation Analysis. It is against the background of limitations to quantitative technical review ratings that counts of citations to publications take on an appeal among some research evaluators. Again, a theoretical framework underpins the use of these counts. The same norms of science that call on researchers to provide public access to their results in the form of publication are thought also to demand that those who receive the results repay the originator with citations. Citations from one paper to another are, in this view, a form of intellectual debt-paying.[61] Whether or not one believes this argument in its entirety, it is clear that the conventions of scientific writing indicate that citations should show some relationship of use or dependence between one article and another.

In this understanding, citations are taken to be in essence an unobtrusive form of wide-scale peer review. Certainly, they add some information to a pure publication count, by indicating whether the work represented in the publications is attracting attention from others in the field. Citation analysts carefully use the word *impact*, rather than *quality*, to refer to what citations count, but they point out that citation counts have been shown in many instances to correlate with peer judgments of quality. One study at the level of individual articles, for example, found that citation counts predicted (in the statistical sense) the quality ratings of each of two technical experts better than the experts' ratings predicted each other.[62] For individual scientists, peer ratings showed correlations with citations in the .6 to .9 range in psychology, physics, and chemistry, although the correlation dipped as low as .20 in sociology, in a set of studies reviewed in the classic volume *Evaluative Bibliometrics*.[63] At the level of university departments, in biology, physics, chemistry, and mathematics, peer rankings and citations showed .67-.69 correlations.[64]

However, a litany of objections has been voiced over the years to equating high citation counts to scientific quality.

---

[61]R. K. Merton, *The Sociology of Science: Theoretical and Empirical Investigations*, Chicago: University of Chicago Press, 1973; W. O. Hagstrom, *The Scientific Community*, New York: Basic Books, 1965; N. Storer, *The Social System of Science*, New York: Holt, Rinehart and Winston, 1966.

[62]Julie A. Virgo, "A Statistical Procedure for Evaluating the Importance of Scientific Papers," *The Library Quarterly*, Vol. 47, No. 4, 1977, pp. 415-430.

[63]Op. cit., Chapter V, pp. 82-121.

[64]Warren O. Hagstrom, "Inputs, Outputs, and the Prestige of University Science Departments," *Sociology of Education*, Vol. 44, Fall 1971, pp. 375-397.

- A small share of citations are negative. Studies in the 1970s showed that the share was negligible,[65] but high levels of citation to the disputed cold fusion results have raised fears again that locally, the influence of negative citations could be strong.
- Citation numbers are highly dependent on field of science, much more so than publication counts. Biochemists, for example, use an average of about thirty references per article, while mathematicians use only about ten. This effect can be normalized in some kinds of analysis, but doing so takes away the simple, intuitive interpretation of citation statistics.
- In many fields, experimental work tends to be cited more frequently than theoretical, and occasional methods papers achieve extremely high levels of perfunctory citation. Citation counts may thus under-value growth in understanding and over-value sheer experimental activity--just the opposite of what one would hope for them as a measure of knowledge quality.
- Because the *Science Citation Index* includes references only from journals, in fields where books are a major publication outlet (including the social sciences), citations undercount even impact.

The differences in citation patterns in different fields of science rule out their use as aggregate performance indicators if any comparison across fields is to be done--for example, if NSF were required to report performance for each of its seven research directorates. Within a field, however, since the limitations are likely to apply with equal force over time, citation counts may be useful for setting baselines of visibility for aggregates of publications. Comparison groups can also be constructed for any aggregate of publications based on matched journal sets, to show where that set of publications stand in comparison with others in the same fields. These are kinds of information that technical ratings cannot provide. The NSF directorates, for example, might determine their baseline citation rates and set as a performance goal to keep fluctuations from those rates within certain limits, or to stay 25 percent above the average citations for articles in the same journals. Since citations peak two to three years after publication, citation information may lag the award of grants by only five to six years, much less than the lag for true outcome indicators. Potentially, then, they could provide useful information for research management purposes, and serve as one among several performance indicators.

Mixed Methods. The strengths and weaknesses of peer review and citation counts appear to be complementary, and evaluators generally advocate using the two

[65]D. E. Chubin and S. D. Moitra, "Content Analysis of References: Adjunct or Alternative to Citation Counting?" *Social Studies of Science*, Vol. 5, 1975, pp. 423-441.

together for detailed program evaluation purposes. A technical review panel's judgments, for example, can be challenged by requiring it to study and respond to literature-based data on the program being evaluated.[66] Conversely, professional evaluators can incorporate both citation measures and peer ratings into an overall evaluation report.[67] These combinations have many advantages for program evaluation purposes, where dialog is possible in the evaluation process.

For summary performance reports as well, a consensus seems to be emerging that descriptive information can fruitfully be combined with performance indicators. As mentioned earlier, the interagency group developing guidelines for research performance indicators agrees that indicators should always be embedded in a narrative, to avoid the worst problems of misinterpretation. Reference to technical review processes within a performance report also seem appropriate, as long as this is not used to sidestep the responsibility of developing some meaningful indicators. If a goal cannot be quantified but the agency checks its performance on it carefully and regularly through technical review, it is perhaps more appropriate to call attention descriptively to the goal and how it is judged, rather than to discard it.

Importance of Research Results

Program managers and participants often perceive the most important characteristics of the knowledge produced by research programs in terms of factors that go beyond both quantity and quality. In disciplinary programs, the theoretical significance of the knowledge is frequently the paramount consideration: Have the researchers in the program enriched the whole field through their insights? Have they developed concepts, methods, or models that apply widely? In mission agencies, a prime consideration is the relevance of the knowledge produced to the practical goal of the program. We refer to both theoretical significance and mission relevance together in this section as *importance*.

User Evaluations. From the standpoint of program evaluation, the key question in judging importance is who does the assessing. Next-stage users are often involved in this judgment. When importance is judged with regard to bodies of scientific knowledge, researchers must judge that quality—but not the researchers supported by the program, nor those who chose the projects it supported. Instead, the next-stage users in this case are researchers outside the program, in the areas where the program's work is claimed to have

[66]Joe Anderson, "New Approaches to Evaluation in U.K. Funding Agencies," SPSG Concept Paper No. 9, Science Policy Support Group, London, October 1989.

[67]The Mitre Corporation, "Evaluative Study of the Materials Research Laboratory Program," MTR 7764, September 1978.

an impact. Agencies that create generic knowledge resources and human capital, as discussed earlier, can in addition identify stakeholder groups for the resources they produce--that is, groups that use the bodies of knowledge and talent pools that the agencies develop, although not the immediate knowledge outputs of specific projects. Such groups can be involved in detailed program evaluation processes.

In mission-oriented programs, next-stage users work in the areas of practice where the knowledge is intended to be useful. Thus, it is quite common to find industrial representatives on evaluation teams; ONR involves DOD technology transfer agents; and the Agricultural Research Service invites large farmers to its evaluation workshops. ARL even included end users--the soldiers who would work with the weapons being developed --in its strategic planning process, opening the door to the inclusion of other end users in research management processes elsewhere.

The state of the art in research program evaluation has not developed effective ways to translate the descriptive knowledge that users bring to the program evaluation process into performance indicators. Nor has it needed to, since users could be involved alongside technical reviewers in any fully developed program evaluation. For GPRA purposes, however, next-stage users may need to be treated as the "customers" for a research program and surveyed for their satisfaction. This would be a step toward evaluating the results, rather than merely the activity, of research. Appropriate survey instruments and samples could undoubtedly be developed. The Army Research Laboratory, for example, includes customer satisfaction ratings in its summary performance indicators, gathering them on a simple customer feedback form sent out with all final project results.

It is well to keep in mind, however, that there are conflict of interest problems in user ratings of research programs. Next-stage users are the recipients of a free service provided by the federal government, and have a stake in expressing high satisfaction with the programs that benefit them, without regard to their efficiency.

Literature-based Tools. Some sophisticated literature-based techniques have been proposed to give strategic overviews and provide background information for judgments of the strategic contributions of program participants.[68] Even advocates do not claim, however, that such techniques can be used independently, without interpretation by technical experts; and they are in fact so complex that they have rarely been used in

[68]See M. Callon, J. Law, and A. Rip (eds.), *Mapping the Dynamics of Science and Technology*, London: Macmillan, 1986; H. Small and E. Garfield, "The Geography of Science: Disciplinary and National Mappings," *Journal of Information Science*, Vol. 11, 1985, pp. 147-159.

practice.[69] No simple GPRA performance indicators based on these methods suggest themselves.

Contributory Goals

Research programs are also generally expected to contribute to certain broad, federal goals, even when these are not listed as among the program's specific objectives. Indicators of performance in relation to these goals should also be included for program evaluation purposes, unless they are not applicable in a particular case (for example, undergraduate training goals in relation to a national laboratory's research program). In theory, indicators on these criteria could be included in GPRA performance plans and reports as well. Examples of such indicators appear in Table 2.2. All the items on the list represent output indicators, which could be gathered from principal investigators at the completion of research projects and aggregated at agency level. Partnership indicators can also be gathered from the published literature. If such data were collected in final project reports, however, it would be important to communicate the portfolio concept clearly to both investigators and program managers. When the projects are gathered into a portfolio, not every project needs to produce each of the outputs on the list, even though in the aggregate, fundamental science agencies expect to create desired outcomes through these routes. To convey any other message would be to limit the flexibility needed for creative work

Table 2.2

Illustrative Indicators for Contributory Goals

Graduate training

- Doctorates earned with support from the program
- Number of recent Ph.D.s who have entered research careers
- Number of recent Ph.D.s who have entered professional practice.
- Professional training provided to non-Ph.D.s by researchers in the program.

Undergraduate training and informal science education

- Undergraduates involved in the research
- Undergraduate course enhancement based on research results
- Informal science education activities
- Popular diffusion of research results

[69]Peter Healey, Harry Rothman, and Paul K. Hoch, "An Experiment in Science Mapping for Research Planning," *Research Policy*, Vol. 15, 1986, pp. 233-251.

Partnerships

Cross-disciplinary partnerships

Cross-sectoral partnerships

Cross-national partnerships

CAUTIONS, CAVEATS AND NEEDS FOR FURTHER RESEARCH

From this discussion, it should be abundantly clear that the methods available for examining the results of research programs may be quite reasonable to use in the context of program evaluation, where multiple indicators are the rule and knowledgeable people are available to integrate them wisely into an assessment. Cautions and caveats about such use have been discussed in the preceding subsections, and are already embodied in the practice of research program evaluation, particularly in the use of multiple indicators and their combination with technical review. In responding to the GPRA requirement for a few summary indicators, however, agencies will apparently need to pare down this full data set to its essential elements. A different set of cautions applies in this situation.

A frequently voiced fear about GPRA is that it will encourage agencies to measure what is easy and neglect what is important. One can easily picture the indicators that would fill this description and satisfy possible administrative requirements for a limited, objective, quantitative set, with which one could set baselines and compare later performance:

- publication counts (year of review)
- citations per publication (lagged three years; compared with average for journals where they were published)
- doctorates produced
    - entering research careers
    - entering careers in practice
- undergraduates involved
- user involvement and satisfaction ratings (in-science users for some programs, outside-science users for others)

The problem with the set, of course, is that it leaves out virtually all of what researchers themselves find important about their work. One could have a government full of programs that performed beautifully according to these indicators, and still be at the trailing edge of every scientific frontier.

The key to responding intelligently to GPRA may therefore lie not in the indicators themselves, but in the larger effort in program evaluation in which they are

embedded and which the Act requires. The indicators, preferably reported in context as the Swedish example above illustrates, can provide a bare-bones description of whether the program is producing the basic expected outputs, and can point toward programs that are particularly in need of evaluation. But the more detailed information that is needed for general program planning and resource allocation, including descriptive judgments and analysis, still needs to come from the more intensive and interactive process of detailed program evaluation.

Many of the key issues with regard to implementation of GPRA, however, lie outside the control of agencies, and in the hands of those who receive and use the performance measures. Optimists about GPRA claim that it will revolutionize government management by focusing program attention intelligently and diligently on results. Pessimists fear that it will create busy work number-generating, then put a simple-minded tool in the hands of decision makers who already pay too little attention to the programs they expand, cut, and re-arrange.

In a recent speech, Senator Roth, GPRA's co-sponsor, said:

> Imagine what you could do if you combined the kind of program performance information envisioned by GPRA with ... program cost-accounting information. We could track the cost-per-unit of activity, and the results of the activity. ... We could have a sophisticated pay-for-performance system that said, "If you achieve all of your program's managerial goals, and do it under-budget, you will get a significant bonus out of the savings you have created."[70]

Where the actual result falls--probably somewhere between the extremes the optimists and pessimists describe--will depend first on what the Office of Management and Budget encourages and requires of agencies as it collates their responses into government-wide performance plans and reports, and second on how the indicators are used in Congress. The first set of results will not be in Congressional hands until March, 2000. If the election trends of the early 1990s continue, most members of that future Congress have not yet been elected, and therefore probably have not yet begun thinking about how they will react to the indicators the research community is now beginning to prepare for their perusal.

[70] Roth, March, 1995, op. cit., p. 6.