Quality of Research in Science: Methods for Postperformance Evaluation in the National Science Foundation (1982)

Pages
129

Size
5 x 9

ISBN
0309327539

Subcommittee on Postperformance Evaluation of Research; Committee on Science, Engineering, and Public Policy; National Research Council

OR

The Quality of Research in Science

Methods for Postperformance Evaluation
in the
National Science Foundation

Report of the Subcommittee on
Postperformance Evaluation of Research
to the
Committee on Science, Engineering, and Public Policy

National Academy of Sciences
National Academy of Engineering
Institute of Medicine

NATIONAL ACADEMY PRESS
Washington, D.C. 1982

The National Academy of Sciences was established in 1863 by Act of Congress as a private, nonprofit, self-governing membership corporation for the furtherance of science and technology for the general welfare. The terms of its charter require the National Academy of Sciences to advise the federal government upon request within its fields of competence. Under this corporate charter, the National Academy of Engineering and the Institute of Medicine were established in 1964 and 1970, respectively.

The Committee on Science, Engineering, and Public Policy was organized in 1981 from the Committee on Science and Public Policy of the National Academy of Sciences. COSEPUP differs from its predecessor committee in representing the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine by including members of the councils of all three bodies.

The study and this report of the work were supported by Contract No. EVL-8115789 between the National Science Foundation and the National Academy of Sciences.

NATIONAL ACADEMY OF SCIENCES
NATIONAL ACADEMY OF ENGINEERING INSTITUTE OF MEDICINE

COMMITTEE ON SCIENCE, ENGINEERING, AND PUBLIC POLICY

2101 CONSTITUTION AVENUE
WASHINGTON, D.C. 20418
202/334-2424

March 8, 1982

The Honorable John B. Slaughter
Director, National Science Foundation
Washington, DC 20550

Dear Dr. Slaughter:

It is my privilege to submit the report prepared by our Subcommittee on Postperformance Evaluation of Research at the request of the National Science Foundation in response to concerns expressed by the U.S. Congress about methods of assessing the results of basic research in science.

There are many approaches to advancing excellence and maintaining accountability of research, including especially the peer review system of evaluating proposals. The objective of our subcommittee's exploration of this matter was to identify ways of evaluating the outcomes of research supported by the NSF. As the accompanying report points out, the outcomes of basic research are, foremost, contributions to the knowledge base of science. Additionally, basic research promotes the maintenance of scientific excellence by strengthening institutional capabilities and by developing a continuing supply of capable young researchers. Postperformance evaluation is one of the means by which the NSF and the Congress can assure themselves and the public that these results are achieved.

Our subcommittee recognizes that any assessment of basic research is bound to be complex and subjective and, thus, controversial. All methods reviewed have significant limitations. Even so, if applied thoughtfully, such evaluation would serve a useful purpose. Thus, after reviewing the subcommittee's report, we have come to the following conclusions and recommendations:

Conclusion 1

Postperformance evaluation can and should be carried out by NSF at the program or division level, that is, through evaluation of aggregations of individual grants or projects.

Accordingly, we recommend that the NSF should expand and strengthen the use of external committees, consisting of persons from a variety of disciplines and experiences, to exercise critical evaluation of the agency's programs. The results of these reviews, conducted periodically, should be reported by the external committees directly to the Office of the Director.

In addition, we recommend that occasional retrospective examinations of specific fields of science should be conducted to trace the origins of significant contributions. Such studies should determine, in particular, the role played by the NSF in advancing the field.

We recommend, also, that the results of these program reviews and retrospective analyses should be used by the NSF in setting research priorities and in allocating resources.

Conclusion 2

Postperformance evaluation at the individual project level is best done in the course of reviewing proposals for renewal of research grants. In this way all of the factors affecting a particular investigation--the quality of the work, the direction the research has taken, the risks involved, and the results obtained--can be taken into account. The subcommittee's survey of NSF's Chemistry Division finds that, to a large extent, this is already being done in that division.

To ensure systematic NSF postperformance evaluation of projects, we recommend that the NSF require throughout all its research divisions that past performance under NSF support be reviewed explicitly in the course of evaluating proposals for renewal of research grants.

Finally, it is the committee's view that the NSF should take the steps necessary to carry out these recommendations and the more detailed activities listed in the last section of the report. The NSF may find it useful to ask an independent group to assess, in a year or so, to what extent the agency has been successful in developing broadly its postperformance evaluation activities along the lines recommended here.

Sincerely,

George M. Low

George M. Low
Chairman

NATIONAL ACADEMY OF SCIENCES
NATIONAL ACADEMY OF ENGINEERING INSTITUTE OF MEDICINE

COMMITTEE ON SCIENCE, ENGINEERING, AND PUBLIC POLICY

2101 CONSTITUTION AVENUE
WASHINGTON, D.C. 20418

SUBCOMMITTEE ON POSTPERFORMANCE EVALUATION OF RESEARCH

March 8, 1982

Mr. George M. Low, Chairman
Committee on Science, Engineering
and Public Policy
National Academy of Sciences
Washington, DC 20418

Dear Mr. Low:

Herewith is the report of the Subcommittee on Postperformance Evaluation of Research. This concludes our exploratory study of methods that might be used to evaluate the results of basic research. The document represents the "interim" report (to be available for hearings in early 1982) called for in the Senate request that led the National Science Foundation to contract with the Academy for the exploratory study.

We were asked to recommend methods for assessing the quality of research produced as a result of NSF support, and to suggest ways in which such assessments could help the agency improve its support of basic research. On the basis of our examination of the procedures of NSF's Chemistry Division, of methods of postperformance evaluation used by industry, private foundations, scientific journals, and other federal agencies, and of other methods employed in the field, we find that

- Postperformance evaluation already occurs at NSF at the project level and at the program level, at least within the NSF Chemistry Division.
- NSF fails to make clear to Congress the degree to which its operating procedures incorporate evaluations of past performance.

- No additional methods of postperformance evaluation that we know of will significantly improve the selection of individual projects.
- Additional or improved strategies of postperformance evaluation should concentrate on aggregate--for example, program--levels and on such issues as the allocation of resources among the subfields of a discipline and the support of young researchers.

The last section of our report recommends seven further activities. We suggest that these activities could be carried out by NSF itself, perhaps assisted through consultants, grants, or contracts. If, however, NSF or the Academies prefer that the subcommittee carry out the first two activities--(1) examining two additional NSF divisions and (2) analyzing information pertinent to postperformance evaluation already available at NSF--I believe it would do so, since that was originally expected of us. Alternatively, the subcommittee would be willing, I feel sure, to advise and consult with NSF as it carries out the additonal activities that we recommend.

Sincerely yours,

W. Allen Wallis
Chairman

SUBCOMMITTEE ON POSTPERFORMANCE EVALUATION OF RESEARCH

W. ALLEN WALLIS (Chairman), Chancellor, University of Rochester
ROBERT F. BORUCH, Director, Program on Methodology and Evaluation Research, Northwestern University
GEORGE E.P. BOX, Bilas Research Professor of Statistics, University of Wisconsin
JOHN M. DEUTCH, Dean of Science, Massachusetts Institute of Technology
CHARLES R. PLOTT, Professor of Economics, Division of Humanities and Social Sciences, California Institute of Technology
W. ANN KING REYNOLDS, Provost, Ohio State University
HOWARD E. SIMMONS, Jr., Director, Central Research and Development Department, E.I. du Pont de Nemours & Company, Inc.
ARTHUR L. SINGER, Jr., Vice President, Alfred P. Sloan Foundation
GERALD TAPE, Special Assistant to the President, Associated Universities, Inc.
EDEL WASSERMAN (Liaison, NRC Committee on Chemical Science), Member, Research Staff, Central Research and Development Department, E.I. du Pont de Nemours & Company, Inc.

SENTA A. RAIZEN, Study Director
LAWRENCE S. WOLFARTH, Consultant

PREFACE

This report presents the conclusions of an exploratory study of methods in the evaluation of basic research in science. The study was motivated by Congressional interest during the past few years in the way the National Science Foundation (NSF) assesses the results of the research the agency has supported.

In 1980, the Senate Committee on Appropriations directed NSF "to secure an independent third party to develop a methodology for post-performance evaluation of scientific research endeavors. . . . This study should strive to identify the criteria and the procedures required for implementing a successful post-performance evaluation effort" (U.S. Senate 1980b). At the end of October 1980, NSF asked the National Academy of Sciences to undertake an exploratory study of methodology for postperformance evaluation of basic research. Early in 1981 the Academy agreed to do so.

The study was originally intended to be carried out in two phases. The first or exploratory phase was designed (i) to review past and current attempts to evaluate completed research; (ii) to assess promising approaches to evaluation, using NSF's chemistry program as the initial substantive field for examining the feasibility and utility of alternative methods; and (iii) ultimately to formulate the rationale and plan for a more detailed analysis of some potentially useful approaches to post-performance evaluation.

To carry out the study, the Academy convened the Subcommittee on Postperformance Evaluation of Research, under the aegis of its Committee on Science and Public

Policy, now the Committee on Science, Engineering, and Public Policy (COSEPUP). The subcommittee is composed of scientists drawn from the natural sciences (chemistry, physics, and biology), the social sciences (economics and psychology), and statistics. The members have had extensive experience in managing research programs in universities, industry, government, and foundations.

At its first meeting, July 28-30, 1981, the subcommittee invited NSF officials and Congressional staff to discuss their expectations of the study. During the course of the study, it asked executives responsible for the management of research in private industry how they assess basic research carried out in their laboratories; their responses are summarized in Appendix A. Evaluation practices of several private foundations were also surveyed. In addition, because editors of scientific journals are called upon to judge the quality of science reported in papers submitted for publication, information was solicited from about a dozen editors about their refereeing and selection procedures. A summary and analysis of the information and comments provided by the editors are given in Appendix B. Along with such specially solicited information, the subcommittee reviewed earlier postresearch evaluation studies of two types: formal assessments of a variety of federally sponsored research programs and examinations of some of the methods employed to evaluate research. Appendix C summarizes pertinent studies in both categories.

Any new evaluation procedures will require expenditures of money, human resources, and time, placing additional burdens on the agency and on the scientific enterprise. With this in mind, we considered it important to find out how much relevant information is already available within NSF and what use is being made of it. Accordingly, several members of our subcommittee spent a day at the NSF Chemistry Division, which had been selected by NSF for the initial analysis of evaluation methods. The subcommittee members reviewed awards, declinations, renewals, borderline cases, proposals from

young investigators, and experimental renewal procedures based on research performance during preceding grant periods. The members were briefed on the Division's procedures and on the activities of the Chemistry Advisory Committee, and they discussed with the staff the data available on the performance of the Division. During the visit, the evaluation responsibilities and studies carried out by the Office of Audit and Oversight were also summarized for the subcommittee members. The results of the visit are summarized in Appendix D.

Our findings as to how postperformance evaluation of basic research should be approached are based on these various sources of information and, more importantly, on our experience as researchers and research administrators. This report, then, concludes the exploratory phase and provides the rationale and recommendations for some follow-on activities. The report has undergone careful review by the subcommittee's parent group, the Committee on Science, Engineering, and Public Policy.

The subcommittee is grateful for the assistance it has received from NSF officials--in particular, Donald N. Langenberg, Deputy Director; Jerome H. Fregeau, Director, Office of Audit and Oversight, and Harry J. Piccariello, Head, Evaluation Staff; and Richard S. Nicholson, Director, Division of Chemistry, and his staff. The subcommittee also expresses its appreciation to Wallace Berger of the Senate Appropriations Subcommittee on the Department of Housing and Urban Development and Independent Agencies, to Helen Gee of the National Institutes of Health, and to the National Research Council's Committee on Chemical Sciences which provided suggestions for a listing of recent chemical advances. Special thanks are due to the industrial managers and to the editors of chemistry journals whose thoughtful responses to our inquiry furnished important documentation. (For names of the industrial managers and the editors who responded to our requests for information, see Appendixes A and B, respectively.)

1 INTRODUCTION AND SUMMARY

> Basic research is performed without thought of practical ends. It results in general knowledge and an understanding of nature and its laws. This general knowledge provides the means of answering a large number of important practical problems, though it may not give a complete specific answer to any one of them. . . .
>
> One of the peculiarities of basic science is the variety of paths which lead to productive advance. Many of the most important discoveries have come as a result of experiments undertaken with very different purposes in mind. Statistically it is certain that important and highly useful discoveries will result from some fraction of the undertakings in basic science; but the results of any one particular investigation cannot be predicted with accuracy.

So wrote Vannevar Bush in his 1945 report, Science --The Endless Frontier. In it, he argued persuasively for the federal government to assume a major role in the support of basic science.

During the Second World War, government support of research and development had expanded from $69 million in 1940 to nearly $1 billion in 1945, largely allocated by an unprecedented and unorthodox office, under Bush, for mobilizing science and technology in the war effort.

On the basis of this experience, Bush proposed a national research foundation that would promote and develop scientific research in the spirit of free inquiry while, at the same time, it would be responsible to the president and to the Congress for its programs. Soon after the Bush report was received by the president and his advisers, legislation was introduced to establish the proposed research foundation. Still, five years elapsed, punctuated by a presidential veto, before the National Science Foundation (NSF) was established in 1950.

The creation of NSF implied a commitment to support pure science as the scientists themselves were accustomed to conducting it. Bush himself, despite his plea to enlarge basic science greatly, had recognized that, in peace time, increased federal support for research in universities and other nongovernmental organizations might engender conflicts between scientists, whose effectiveness benefits from independence, and legislators, who are accountable for the use of public funds. It became inevitable that in time, as NSF grew and became progressively more important to the performance of scientific research, difficulties would be encountered in continuing the "bargain by which scientists would get support for basic research which government officials would hope would lead to applied research and to useful developments" (Price 1978).

Federal expenditures for basic science increased significantly and consistently until the late 1960's, growing from $715 million in 1960 to $2.3 billion in 1968. Since then, growth in current dollars has continued to nearly $6 billion in 1981, but in constant dollars annual funding has remained nearly level for the last five years (National Science Board 1981). (In terms of constant dollars, funding nearly tripled between 1960 and 1968, but has increased by only 7 percent since then.) The federal government's share also has increased considerably. In 1940, it paid for only one-fifth of the nation's basic research; today, it pays for slightly more than two-thirds, with half of the funds supporting work in the nation's universities. NSF has been a key factor in this growth. From 1963 to 1981, its support of basic research grew from $141 million to $877 million in current dollars, or more than doubled in constant dollars.

The National Science Board (1981), which governs the operation of NSF, recently called attention to the new tensions arising in the performer-patron relationship,

which go beyond the debates about the adequacy of current or proposed funding levels to the output of basic science. "Many of the concerns regarding administrative requirements and their impact on research performance stem from increased pressures for greater accountability for public resources," the Board states. In the past few years the concerns have become manifest in extensive time-keeping and cost-accounting procedures for university researchers and inquiries into the peer review systems for awarding research grants (see Cole et al. 1978, Staats 1979, National Commission on Research 1980, Mac Lane 1980, Wilson 1980, Cole et al. 1981, and GAO 1981).

At the heart of the problem lies the fact that it is difficult, and often misleading, to evaluate the outcomes of basic research in the short term. Yet, the Congress expects to know from year to year what it has been "buying," so that it can decide whether and at what levels to continue such support. Any use of public money obligates the user to account for it.

Since 1978, the Senate Committee on Appropriations has been raising questions about postperformance evaluation of basic research during the annual NSF appropriations hearings. In 1979, the Committee requested NSF to develop "a coherent and effective system of postresearch evaluation" (U.S. Senate 1979). There was an important change in direction the following year. On the basis of discussions with scientists and NSF's mixed experience with some exploratory evaluations, the Senate Committee did not call--as it had earlier--for the implementation of an evaluation system, but emphasized the need to develop better methods for appraising NSF's research functions before an evaluation system was designed and used. The request in 1980 makes clear the Senate Committee's intent that methods developed for postperformance evaluation serve two purposes: first, provide a continuing accounting of the outcomes of the research supported by NSF, and, second, identify the factors that determine the productivity of basic research in order to improve the funding decisions made by the agency. In the Senate's own words:

> The Committee remains convinced that postresearch evaluation efforts are meaningful and important. The Committee feels that post-performance evaluation activities will provide the scientific community and the NSF

> with a better understanding of how successful science is supported and performed. An ongoing examination of selected projects and research programs is necessary in order to provide a clearer insight into what criteria are necessary for the most productive research. There are many subtle variables in this complex equation--the individual researcher, the various strengths of his or her institution, the originality of the idea, the contribution to knowledge the research makes, and others. The Committee feels the need to determine which factors are of primary importance.
>
> As a result, the Committee directs NSF to secure an independent third party to develop a methodology for post-performance evaluation of scientific research endeavors. . . . This study should strive to identify the criteria and the procedures required for implementing a successful postperformance evaluation effort (U.S. Senate 1980b).

From the point of view of NSF, it is clear what should result from the research the agency supports: gaining scientific knowledge, training future researchers, increasing research capacity, and encouraging scientific collaboration. But how to measure the results is not so clear. As Vannevar Bush observed, scientific research pays off in the aggregate, but the contributions of individual projects are indeterminate. Basic research entails risk of failure, and some payoffs may be far in the future. These conditions make it difficult, if not impossible, to measure the outcomes of basic research objectively, accurately, or with certainty, especially in the short term. NSF, accordingly, faces a dilemma: While it has been directed specifically to evaluate scientific performance, it understands that the task is not only complex but, unless done with great care, may inhibit rather than promote the quality of basic research.

Concerns for accountability should concentrate, in the first instance, on the NSF peer review system for evaluating proposals, which guides the agency's core activity, that is, making decisions about what proposed research projects to fund. A recent five-year National Academy of Sciences study (Cole et al. 1978, 1981) found

the system to be operating fairly, with grants being awarded in accordance with the best judgments of the scientific community as to the scientific promise of the work to be performed. Nonetheless, accountability also requires serious attention to the quality of completed work. Chapter 2 of our report cautions, however, that criteria for evaluating research outputs must not inadvertently discourage the intellectual creativity and risk-taking that are a sine qua non for achieving major advances in scientific knowledge. Such criteria must recognize that uncertainty properly characterizes the typical proposal, that there are multiple outcomes of research, that both positive and negative results are useful, and that there may not be agreement about the significance of new knowledge. Measures of value must also recognize that research fosters important benefits in the training of the next generation of scientists and the development of institutions of excellence.

All of this suggests the difficulty of identifying hard and fast predictors or indices of successful basic research, and hence, of defining criteria for measuring its outcomes. Furthermore, to the extent that one seeks clues in the assessment of results for improving future proposal review, the problem is compounded by the fact that results are sometimes not appreciated until many years have passed.

Having set forth the problems associated with evaluating science in Chapter 2, we turn, in Chapter 3, to a consideration of some methods of evaluation and the promise they may offer as useful and reliable tools for postresearch evaluation by NSF.

The most important method is peer judgment. Peer judgments permeate the scientific endeavor. They are the foundation of virtually all decisions that affect who will learn, who will teach, who will advance, who will perform research, and what results will be published and later used. At NSF, peer advice is used to evaluate proposals and to oversee the functioning of the research support units. We found that, at least in the Chemistry Division, a very high proportion of NSF-supported investigators submit renewal proposals to continue their research efforts. It follows that, to the extent that the performer and the performance under prior grants is assessed by the reviewers of renewal proposals, the peer review system provides a built-in means for the postperformance evaluation of projects. (Cole et al. 1978, did

not find a high correlation between grants awarded and an investigator's "track record" as measured by publication and citation counts. However, the Cole brothers and their co-workers used data from proposals reviewed in 1974; since then, NSF guidelines have put increased emphasis on the importance of an investigator's recent performance.) Beyond information derived from the review process, evaluative information pertinent to postperformance accountability is also available from the peer advisory committees that oversee division procedures and from NSF's other analytical and reporting activities.

Evaluation methods commonly employed to assess the results of basic research in other organizations, both in government and in industry, vary from NSF procedures mainly in detail. Peer judgment is the key method used. Other methods are used from time to time to corroborate the assessments made through the use of peer ratings.

One such method, bibliometric analysis, involves the counting of articles published by researchers and the number of subsequent citations to their work by other scientists--an approach that aims to summarize qualitative peer judgments in a quantitative way. Use of the technique as a supplementary tool for evaluation has been attempted by NSF to evaluate its oceanography program, and by the Rand Corporation to evaluate biomedical research programs of the National Institutes of Health (NIH). A variety of limitations leaves the utility of such techniques in doubt, particularly for assessing the productivity of individuals. Bibliometric measures are best used when the groups being assessed are large and comparable to each other, and when the measures serve the auxiliary function of strengthening confidence in peer judgment.

Case studies about scientific discoveries have provided interesting insights into the social, historical, and intellectual factors that affect the research process. A few additional case studies may shed further light on how research productivity might be improved, although we doubt that analysis of past scientific activities can be very helpful in assessing proposed research. Since case studies are costly, their use is likely to remain limited. Moreover, case studies are often protracted. Hence, they offer little practical assistance in the short-term evaluation of research outcomes.

Retrospective analyses such as NSF's T.R.A.C.E.S. (for Technology in Retrospect and Critical Events in Science, Illinois Institute of Technology 1968) are a useful means for estimating the contribution of earlier discoveries to significant advances and applications--whether originally foreseen or not. But this technique, like the more intensive case studies, cannot yield assessments of the quality and contribution of research in the short run and thus are not helpful to the program manager in making funding decisions. Retrospective analysis would be useful in assessing the extent to which the research supported by, say, NSF a decade or more ago led to important scientific developments.

An evaluation method worthy of further consideration is prospective analysis--i.e., planning in advance how to evaluate the consequences of the agency's experiments in funding and review procedures. For example, it would be useful to learn to what degree the Chemistry Division's experimental procedures (described below) for the handling of renewal requests improve the effectiveness of its decisions.

As a result of our exploration of the general problem of evaluating research, the suitability of various evaluation methods and strategies, and the procedures now employed by NSF's Chemistry Division, we reached a number of preliminary conclusions, some of which suggest further steps for more detailed consideration of postperformance evaluation. These will be found in Chapter 4.

In the context of this study, we believe that methods for postperformance evaluation must serve two purposes: to improve NSF's judgments in supporting high quality science and to provide a basis for demonstrating the quality of its stewardship of public funds. These two purposes should be served by means that are mutually reinforcing. This can best be accomplished by adapting and using for postperformance evaluation those procedures already employed in the support of research, so as to avoid wasteful and duplicative effort.

The assessment of NSF's performance must proceed from the premise that the key to the agency's success in fostering research of high quality is its management of external peer review of individual research proposals. Moreover, the review of the progress and results of funded research in the course of assessing renewal proposals is the best opportunity available to NSF for

systematic and comprehensive postperformance evaluation of individual projects. We know of no additional method of postperformance evaluation that is likely to improve significantly the process of selecting individual projects for funding.

This conclusion is based on our observations of management and current review procedures of the Chemistry Division. Any follow-up of this exploratory study should establish whether the Chemistry Division is characteristic of the agency as a whole. It also would be valuable to examine ways in which the postperformance evaluations implicit in renewal reviews and the various oversight activities, as well as ongoing management experiments, could be better articulated and more clearly communicated.

We believe that postperformance evaluation could be used more effectively in improving decisions about the allocation of resources among fields and subfields of science. The potential for evaluations of this type at the program and division levels has not yet been adequately realized. Accordingly, we make some recommendations for exploring additional sources of evaluation, possibly with reports going to the director and assistant directors, that will be concerned with questions of allocations, the relative strengths and weaknesses of various fields and subfields of scientific inquiry, and the adequacy of attention paid to other NSF goals, such as fostering the continued vitality and strength of science through the support of creative young researchers. It is on these aspects of NSF performance that a follow-on to this exploratory study should concentrate. We believe that evaluations at the level of allocations to programs and divisions will yield more significant results for NSF management than additional evaluations at the project selection level.

To be most useful, performance evaluation should be integrated into the management structure of NSF. Evaluation activities should involve both the affected program offices and the NSF evaluation and analysis staffs. The necessary condition for effective evaluation is a capable internal staff augmented by external advisory committees and contractors.

2 PROBLEMS IN EVALUATING BASIC RESEARCH

In considering the results of basic research, the standards which should be used in evaluating it, and the uses to which evaluations could be put, serious difficulties arise from certain inherent characteristics of such research: its uncertainties, its multiple consequences, its cumulative nature, and its transferability. Paradoxically, it is these very characteristics that provide the argument for public support of science, because they make basic research a high-risk investment, and it is not likely that private parties can capture the benefits.

Uncertainty in Research

Lewis Thomas (1974) has caught the essence of what distinguishes basic research from applied research:

> Surprise is what makes the difference. When you are organized to apply knowledge, set up targets, produce a usable product, you require a high degree of certainty from the outset. All the facts on which you base protocols must be reasonably hard facts with unambiguous meaning. The challenge is to plan the work and organize the workers so that it will come out precisely as predicted. For this, you need centralized authority, elaborately detailed time schedules, and some sort of reward system based on speed and perfection. But most of all you need

> the intelligible basic facts to begin with, and these must come from basic research. There is no other source.
>
> In basic research, everything is just the opposite. What you need at the outset is a high degree of uncertainty; otherwise it isn't likely to be an important problem. You start with an incomplete roster of facts, characterized by their ambiguity; often the problem consists of discovering the connections between unrelated pieces of information. You must plan experiments on the basis of probability, even bare possibility, rather than certainty. If an experiment turns out precisely as predicted, this can be very nice, but it is only a great event if at the same time it is a surprise. You can measure the quality of the work by the intensity of astonishment. The surprise can be because it did turn out as as predicted (in some lines of research, 1 percent is accepted as a high yield), or it can be confoundment because the prediction was wrong and something totally unexpected turned up, changing the look of the problem and requiring a new kind of protocol. Either way, you win.

The uncertainty attached to doing basic research makes it particularly difficult to predict results or to assign values to them in a common metric. Evidence on the degree to which the outcomes of basic research are unpredictable comes from a retrospective study of significant advances in four disciplines from 1950 to 1976 (Kruytbosch 1978, see also Appendix C). Of the 65 advances examined, 37 (59 percent) resulted from grants for which the proposals did not mention the advance as a specific goal of the research, though for 26 of the 37 the advance was in the same general area as that of the proposed research. It is the difficulty of predicting outcomes that makes inadvisable any method of postperformance evaluation in which individuals awarded research grants are held accountable for achieving the specific goals set forth in the proposal. The General Accounting Office (GAO 1981), in reporting its study of renewal procedures at NSF and NIH, noted that the reviewers and agency administrators it interviewed were not concerned

about the failure of most researchers to accomplish all of the objectives set forth in the original proposals because they recognized that the actual results are more important. Though GAO did recommend to NSF that it require applicants for renewal grants to restate the specific aims and overall objective of the preceding grant, it did so to enable reviewers to determine whether the proposed research had been attempted and not for the purpose of scoring the applicant on progress made toward specific goals.

Not only is there uncertainty about forecasting outcomes in basic research, there is also little agreement on the factors that influence outcomes or on the underlying events that may lead to success or failure. Because basic research is a highly uncertain and poorly understood process, it may be desirable to encourage with public funds those projects whose results are most difficult to predict. Thus, if the number of projects funded that achieve their initial objectives is very high, it may mean that the support strategies were too conservative. The demand that federally supported research projects achieve narrowly stated objectives is not only antithetical to the justification for public spending on them, it also can be counterproductive to the promotion of research of high quality and significance.

Multiple Outcomes of Research

The most obvious result of basic research is some contribution to scientific knowledge, but research may also yield education and training of future researchers, institutional benefits, and increased communication among scientists. Most postperformance evaluations have attempted to assess knowledge-related outcomes. The usual approach has been to judge through peer review the quality of the contribution, which may consist of amended theories, empirical findings, or improved techniques and methods. Another approach has been to link advances in knowledge to economically or socially valuable applications. Since research results usually appear in published form, still another procedure for establishing impact has been to count publications resulting from research efforts or citations to them. A second outcome of research projects, particularly those carried out in universities and other educational settings, is the benefits derived from the training of young scientists,

usually at the graduate or postdoctoral level. As Ziman (1968) points out, young scientists who have opportunities to work closely with established researchers--and the top researchers in particular--learn not only procedures and techniques but also how to identify scientifically important problems and design research programs to solve them. Third, the organization that houses a research project also benefits by increasing its skill in doing and managing research. A project may bring together researchers to work on a problem of common interest; the researchers develop techniques, resources, and substantive knowledge, which, over time, enhance the research capabilities of the organization and of other organizations that the researchers may subsequently join. (See, for example, the account by Edge and Mulkay 1976 of the development of radio astronomy in Britain.) Hence, science as a whole benefits. A fourth consequence of doing research is the effect on networks of communication among scientists (Crane 1972), which may be reinforced or attenuated by the involvement of specific individuals in a particular research project, with subsequent long-term implications for patterns of scientific collaboration.

That research efforts usually have multiple and noncomparable outcomes suggests that no single measure can fully reflect the output of a set of basic science projects. Different methods of evaluation may have to be devised for different purposes. At the very least, it means that evaluation should not focus too narrowly upon one measure of research output in a manner that is detrimental to other beneficial aspects of scientific activity.

A danger we recognize is that measures of performance can become self-fulfilling criteria that researchers attempt to satisfy, leading them to deemphasize other important (but non-evaluated) aspects of their work. For example, it seems likely that using publication counts as the principal measure of productivity creates an incentive for the investigator to publish results as soon as possible, even prematurely, and to produce many short papers, or to neglect the training of graduate students. Such a criterion, if used by a funding agency, might also discourage highly original proposals that are likely to carry a relatively greater risk of failure or that are in fields at the vanguard, where there is less likelihood for many near-term citations.

Sometimes, in the course of scientific research, the results are negative. Negative results can be valuable because they may show that certain modes of research,

certain techniques, or certain hypotheses are not useful. Sir Karl Popper has argued in his book, The Logic of Scientific Discovery (1959), that there are no absolute or proven theories in physics, say, or mathematics, only those that have not yet been disproven (or, as he puts it, "falsified"), suggesting that often the most interesting results are those that demonstrate something to be false which had previously appeared to be true. The benefits of negative results lie in changing the direction of research to other, more promising endeavors, and in saving time, effort, and money by avoiding blind alleys. A famous example of negative results changing the direction of research comes from physics. In 1956 Lee and Yang questioned, at least for weak nuclear interactions, the concept of conservation of parity, which held that nature detected no difference between right and left for the curious behavior of atomic particles. Only a few months after Lee and Yang reached this revolutionary conclusion, three teams of experimenters in different U.S. laboratories showed that, indeed, the "law" of parity did not hold, and in 1957 Lee and Yang shared the Nobel Prize in physics (Morrison 1957).

A notorious example of a "blind alley" was research on "polywater." In the 1960's, Soviet chemists announced that when distilled water vapor was allowed to condense it acquired a polymeric molecular structure with formidable properties: Being superdense, polywater froze and boiled at abnormal temperatures, and some scientists even warned that it might transform ordinary water to jelly and make our planet uninhabitable. Skeptical but nevertheless concerned about the implications, U.S. agencies supported research on polywater by many highly regarded chemists. By the early 1970's, the quest for polywater ebbed as it became clear that polywater was nothing more than dirty water, contaminated by silicon leached from the glass or quartz tubes and pipes used during experiments (Franks 1981).

The Cumulative Nature of Science

"All science is the search for unity in hidden likenesses," wrote Bronowski (1956). "The search may be on a grand scale as in the modern theories which try to link the fields of gravitation to electromagnetism [though] there are discoveries to be made by snatching a small likeness from the air too if it is bold enough."

Scientific research is necessarily connective. Most users of the results of basic research are invariably other scientists. The whole enterprise does not take on a value unless there is a collaborative pattern of research outcomes that can then be collected and applied to a related or more complicated problem. Therefore, if basic research is evaluated project by project as results become available, the apparent "worth" of each project could be zero, even though the value of the whole pattern of activity might be enormous. Contemporary estimates are often difficult to agree on--and sometimes wrong. Sadi Carnot's fundamental paper on thermodynamics, appearing in 1824, was not recognized as important until 1834, and then only by one scientist, Emile Clapeyron; after that it took another decade before Carnot's work was appreciated, largely through William Thomson's research (Holton 1978). Similarly, Yukawa suggested in 1935 that atomic nuclei were held together by "forces" like photons of ordinary electromagnetic forces. The next year Anderson discovered a new subatomic particle, the meson, but it did not interact with atomic nuclei as Yukawa had predicted. It was not until 1947 that Powell found a heavier particle, the pi-meson, which met all of Yukawa's requirements (Yang 1961).

The true "revolutions" in scientific research are few and far between. Much of science consists of testing refinements of theories, providing additional data, exploring new avenues that may not prove productive. But such "normal" science (Kuhn 1970), or the filling in of detail, often provides the basis for research breakthroughs. The implication of this is that routine research--which constitutes the activities of most scientists--requires evaluation in the broader, cumulative context.

Since the ultimate importance of a piece of work may not be understood or appreciated until it can be fitted into a broader corpus of work developed subsequently or until more sophisticated instrumentation becomes available, its potential for opening up new lines of inquiry or for practical application may not be realized for years. Consider the case of the 1981 Nobel prize in chemistry awarded to Fukui and Hoffmann. Fukui first published his frontier molecular orbital theory of chemical reactivity in 1954, at a time when most theoretical chemists doubted that reactivity could be reduced to anything so simple. Not until 1964 when Hoffmann (with Woodward) independently developed and formulated their

molecular orbital theory in such a way that it could be directly utilized by experimentalists did it achieve widespread recognition (Streitwieser 1981). Comroe (1977) gives a number of examples of potential applications in the biomedical area that were long unrecognized, such as the development of sulfa drugs.

Thus, the quality and significance of scientific work cannot always or even usually be estimated with certainty right away. This has implications for postperformance evaluation. The three-year period that has been used in some previous attempts to evaluate research outcomes will often be much too short. It might, in fact, be harmful to use only a three-year period--harmful to the extent that it may encourage agency officials to favor "safe" projects that promise quick, publishable results and to reject "off-beat" or "long-shot" projects that do not fit accepted paradigms in the field.

Transfer of Knowledge

There is one more dimension of basic research that makes it difficult to assign values. Knowledge itself is transferable, and most basic research knowledge is freely transferable within the scientific community. A dramatic example comes from the field of atomic energy. During Christmas 1938, Otto Frisch, a young physicist, visited Sweden to stay with his aunt, Lise Meitner, who had just received a letter from her former colleague, Otto Hahn. From this letter, Frisch learned that Hahn and Strassmann had split the uranium atom by neutron bombardment at Germany's Kaiser Wilhelm Institute of Chemistry. When Frisch returned to his laboratory in Denmark, he informed Niels Bohr, who was embarking for the United States to speak to the American Physical Society. So, by a chance series of circumstances, U.S. scientists learned of Hahn's work and its explosive implications in January 1939. Within days of Bohr's address, Hahn's experiment was repeated at Columbia University, the Johns Hopkins University, and the University of California (Clark 1961).

Thus, even before knowledge enters the public domain, the practitioners of a specialty are often informally made aware of new discoveries, methods, and data, in order that they might evaluate the significance. Bound together by shared interests and goals, they ensure the accuracy and quality of what eventually appears in the journals; in return, they may utilize the information

to guide the course of their own research with or without public acknowledgement of the influence. Knowledge transmitted through private and informal channels cannot be traced easily.

Once knowledge enters the public domain through publication in a scientific journal, there is no way of knowing for certain who uses it or how it is being used, except for specific references in subsequent work. The indirect impact on the thought and imagination of other scientists is not easily established. In consequence, it is often difficult for anyone outside the specific community to determine the implications of new knowledge about theories, processes, and techniques. Particularly in the case of negative results, if someone has discovered that a technique will not work or that a chemical process is not feasible without expensive apparatus, and such knowledge enters the public domain, the benefits of that knowledge accrue to those who no longer spend time on marginal lines of research. Since the knowledge is available to all, its use cannot be traced and its benefits cannot be indexed.

Improving the Process

Given some of the characteristics of basic research just discussed, the expectation that certain --especially quantitative--techniques for postperformance evaluation can be used to improve significantly the process of public support of research and increase scientific productivity is likely to be unfulfilled. The expectation is based on a mechanistic input-output model, which largely sets aside the context that surrounds research. Such a model can produce information on the outcomes of a process, provided these can be assigned values, and valid and reliable measures for the values are available. But as noted, neither of these provisions holds for basic research. Moreover, such input-output models do not illuminate the research process. The measurement of output alone (even if it could be accomplished) will hardly add much to existing notions about factors that tend to make for success, such as the track record of the individual, institutional capability, originality of proposed research, and soundness of method.

Because of the uncertainty surrounding the research process, models do not exist that are sufficiently detailed

to allow testable inferences about how scientific productivity might be increased. In summarizing the available studies on the process of research and development (R&D), Plott (1974) notes: "It is our opinion that in the areas we have reviewed [including R&D management, structure of decision making in R&D, cost-benefit and production functions, and screening and committee processes] there is a great need for basic theoretical and experimental work. . . . The preponderance of written works provide anecdotes and ad hoc theories. There exists a plethora of opinions but the instances of integrated theories, replicable results and precisely formulated models are very sparse indeed." Bringing all this together would entail a research program beyond the capacity of our subcommittee. And, as has been amply demonstrated in the attempts to evaluate other types of complex human activity--teaching and learning, for example--unless the process is understood, evaluation of outcomes produces little that is useful in making such a process more productive.

3 METHODS OF EVALUATION

Assessing research requires an understanding of the inner logic of what is going on in any particular piece of research and how it fits into a larger pattern within a field or specialty. Therefore, valid methods for judging research outcomes depend, either directly or indirectly, on the judgments of other scientists who are active in the field. In this section, we describe the various objective and subjective methods that have been used in deciding on the value of scientific work. While the Congressional request deals with the evaluation of research after it is completed, i.e., postperformance evaluation, the methods and criteria appropriate for this purpose are related to the evaluation of proposed research. In each case, past research performance and results are important; relationship of proposed or completed work to other work in the field is another criterion. In particular, the critical appraisal of peers serves as an evaluation of both newly proposed and completed research.

Peer Judgment

Peer judgment permeates the scientific endeavor. It determines the course of a scientific career--entry into a doctoral program and award of the degree, appointment to a faculty or other professional status, granting of tenure, and ranking within a field. Peer review determines the allocation of funds that will be made to scientists and to areas of research. Results are published

or not on the basis of peer judgments. Even the standing of research institutions depends on the perception of scientific peers.

Scientists are constantly making judgments of the importance and quality of research. The judgments become evident in decisions about continuing or limiting a particular line of scientific work. Through this exercise by scientists of their own authority over each other (Polanyi 1962), science regulates itself. This self-regulating process takes place at many different levels and in many different places, and it involves many different people. In this respect, peer judgment is pluralistic, decentralized, and pervasive.

When explicit judgments are necessary--in editing journals, say, or in appointing or promoting scientists within organizations--formalized processes of peer review are used. Such processes play a large part in the operations of the government agencies that support basic research. Peer review is used to evaluate individual proposals; peers advise on research priorities and programs and often help steer an agency like the NSF.

The Use of Peer Judgment at NSF

Since NSF is an integral part of scientific research in this country, peer advice is central to its operation, management, and staffing. Explicitly, however, there are three internal uses of peer judgment that bear directly on postperformance evaluation: proposal review, especially the review of renewal proposals; advisory committees that oversee each division; and special studies carried out by the evaluation and policy analysis units of NSF. There are also external peer judgments of the research produced with NSF support that are made quite apart from the agency and its advisors.

Proposal Review. The most important function carried out by NSF, the funding of individual research proposals, relies solidly on peer review by outside researchers. Procedures vary within the agency. The most frequently used method is to solicit opinions by mail from scientists who are actively engaged in the area. A standard form and instructions are sent to reviewers, together with the proposal and relevant publications, and they are asked to rate the proposal and write an assessment. Some divisions use panels of experts who meet as a group to make recommendations;

still others use a combination of mail and panel review. NSF program officials have some latitude in making decisions, but peer judgment appears to weigh heavily. Cole et al. (1981) state about the peer-review process at NSF: "There is a high correlation between reviewer ratings and grants made . . . . The scores given proposals by reviewers were the most important factor in funding decisions." Generally speaking, proposals rated as "excellent" (5) or "very good" (4) are awarded funding; those rated lower are not. The fact that proposals are reviewed probably leads to a self-screening by applicants, helping to increase the quality of proposals received by NSF.

A large proportion of proposals received by NSF consists of renewal requests--that is, proposals for work that is to follow research performed under a current grant. NSF program officials estimate that renewals are sought by more than 90 percent of the investigators holding grants (U.S. Senate 1980a). In the Chemistry Division, the percentage is slightly higher (see Appendix D). This division receives between 825 and 850 proposals each year and awards 325 to 350 grants (excluding second- and third-year funding of previously awarded three-year grants); three of every four of these grants are renewals. Thus, insofar as NSF program officials stress previous research achievement as one criterion for judging the quality of all proposals, the peer review of renewal proposals serves as one important means of postperformance evaluation. Kruskal (1975) points out, however, that considering only renewal proposals omits the possibility of evaluating work that resulted from proposals not funded by NSF or work that did not lead to a renewal proposal.

In a recent study of 50 NSF and 25 NIH basic research grants, GAO (1981) criticized NSF on the ground that direct evidence of progress on the preceding grant did not play a sufficient role in the evaluation of renewal proposals (see Appendix C). Similarly, both of the studies by Cole et al. (1978, 1981) found low correlations between NSF reviewers' scores and bibliometrically derived measures of the past productivity of investigators. By contrast--and perhaps because of the recent changes in NSF guidelines emphasizing previous research performance--we found that reviewers more often than not discuss an investigator's record in some detail, although they do not always use the separate space provided for this purpose on review forms. From our

inspection of a number of Chemistry Division folders illustrating different categories of proposals and funding actions (see Appendix D), it is apparent that the declining productivity of an investigator has led on occasion to the rejection of a renewal proposal--even in the case of eminent and formerly productive researchers.

The NSF chemistry programs illustrate some of the difficulty of seeking postperformance evaluation of accomplishments achieved under previous NSF grants. Frequently, chemists who are awarded NSF grants have more than a single source of financial support from, say, NIH or another federal agency. The investigator's performance on any one grant benefits from his total research effort, and reviewers will find it difficult to separate the outcomes that flow from different but related projects. It is likely that most reviewers implicitly evaluate the complete record of recent accomplishments of an applicant rather than focus on the pieces of the research supported by NSF.

NSF is experimenting with renewal procedures that put even more stress on previous productivity. "Accomplishment-based renewal" procedures, an option open to all NSF chemistry grantees, allow the investigator to submit a four-page proposal (instead of the usual 15 pages) accompanied by selected reprints and a list of all publications produced during the preceding grant period. For now, this optional renewal procedure is limited to the Chemistry Division. A second alternative, the "creativity extension," is restricted to 10 percent of grantees eligible for renewals in any one year. Program officials select highly creative and productive grantees who are awarded two-year extensions of their existing three-year grants, without needing to submit renewal proposals. Each of the alternatives can be implemented for only one renewal cycle and must be followed in the next cycle by a standard proposal and an external peer review, if the grantee wants further funding from NSF.

Advisory Committees. Peer advisory committees, generally meeting twice a year, have been used for some time by the NSF divisions to provide advice to the staff on significant developments in the field. Since 1979, the advisory committees also have been charged with reviewing division and program operations in very specific ways (NSF 1979). Each committee is required to report at least every three years on the functioning of the proposal review process; on the balance among

programs within a division as to size and number of awards, subject matter, and age and geographic distribution of principal investigators; and on the question of whether the program is meeting NSF objectives. The most recent report of the Advisory Committee for the Chemistry Division (NSF 1980b) is particularly detailed. It is based on three days of review by some 30 outside experts (see Chemical and Engineering News 1980). According to program officials, recommendations in the report led to some redistribution of funds among the chemistry programs in the Division.

The task of the Advisory Committee was undoubtedly aided by the detailed statistics that the Chemistry Division compiles on its operations. Our selective scanning of less detailed reports by the advisory committees for some other NSF divisions has made us aware that this performance review procedure is highly variable in the amount of information it produces.

Occasional Studies. From time to time, NSF carries out or contracts for special studies that are concerned with outcomes or consequences of research supported by the agency. Most of these studies use peer judgment to evaluate the quality of the work done. Generally, such ad hoc studies are intended to meet a specific request from Congress or from NSF management. Several of these studies are described in the first section of Appendix C, including the evaluation of the oceanography program performed by NSF's Office of Audit and Oversight (NSF 1980a) in response to an earlier Senate request for postperformance evaluation.

Externally Generated Peer Judgment. The products of the research supported by NSF are subject to the same scrutiny and value judgments as all scientific work. Papers by principal investigators are screened by editors and reviewers before they are published in any of the major scientific journals. (See Appendix B for an analysis of this process.) After results become part of the open literature, scientists decide for themselves whether to use the published work, depending on their judgment of its quality and significance. Therefore, publication in refereed journals offers one independent means for assessing the performance of research efforts funded by NSF; the rate of acceptance or rejection of papers with NSF sponsorship can be compared with the rate for all papers submitted to relevant journals (i.e., those that cover areas in which NSF is active) that keep a file of all submitted papers.

The subsequent judgment of peers on the importance of prior work, apart from their use of it in their own research, is specifically elicited in carrying out retrospective studies. Thus, Kruytbosch (1978) asked peer panels to select innovations in four fields and then traced the NSF contribution to each of the innovations. (For results, see Appendix C.) In another study (also described in Appendix C), commercial products that had been awarded prizes in peer-judged competitions were used as a starting point to trace the contribution of NSF-sponsored research to industrial innovation (NSF 1981).

Another independent criterion for assessing the effectiveness of NSF programs can be derived from the standing of graduate programs that are compiled on the basis of peer judgments and publication records (see, for example, Gaurman 1980). The problem with using departmental standings as an aggregate-level measure of the quality of research being done at an institution is the likely failure of such standings to reflect very recent achievements and changes in faculty rosters. Standings from prior decades in fact have been used in the Cole et al. studies (1978, 1981) as one indication of a grant applicant's ties to an "old-boy network." (They concluded that proposals from scientists at major institutions were not treated more favorably by reviewers from major institutions.) Assuming that adjustments are made, standings based on the most currently available information can be employed to determine the extent to which NSF programs have provided support to researchers in the faculties deemed highly productive by their colleagues.

Peer Review in Other Organizations

Peer review serves important evaluative functions in all types of science-related organizations. It is apparent from the summaries that follow that the methods used by industry and by other federal funding agencies for assessing basic research vary from those of NSF mainly in procedural detail.

Performance Evaluation in Industry. Scientific and engineering research is an important part of U.S. industry. For the most part such research is relatively short term and centers on technological objectives. A small number of large corporations, however, conduct

fundamental scientific research that spans chemistry, physics, engineering, and the life sciences and that is essentially indistinguishable in kind from research being carried out at the foremost universities. Indeed, there is considerable movement of researchers back and forth between universities and industrial laboratories engaged in basic research, especially in the field of chemistry.

We surveyed six large industrial laboratories and found that their research managers assess the productivity of basic research efforts through a kind of peer review. Frequently, reliance is placed on academic consultants and visiting committees who advise generally on the quality of staff members and their work. Management is guided by an investigator's record of scientific achievement over a period of years rather than by the success of his latest project. All corporations perform annual in-house performance reviews, and high ratings can come from failures as well as from successes, particularly in cases where creativity has been shown in the conception and execution of a project. (For more detail, see Appendix A.)

The subcommittee has given substantial weight to the experience of industrial executives in evaluating their basic research programs. It is important to recognize that industry does not have methods of performance evaluation different from those of NSF where the tasks are analoguous. Both industry and NSF have found nothing that serves better than some form of peer review for the evaluation of individual research projects.

Performance Evaluation in Other Agencies. Like NSF, NIH uses peer review to assess the merit of proposals submitted to its extramural support programs, though the NIH system entails two sequential levels of review and a more highly structured process that leaves little discretion to program officials. In the case of renewal proposals, evidence must be presented of past performance, including the extent to which objectives of the immediately preceding grant were met and a list of publications that resulted from it.

A special structure of external advisors performs continuous evaluation and guidance of NIH programs. Each of the institutes has its own Board of Scientific Counselors that, twice each year, reviews all ongoing or proposed research on the basis of formal presentations by individual investigators. Similarly, the U.S. Environmental Protection Agency (EPA) also asks funded

investigators to present and defend their findings before an external group of experts. In EPA's procedure, the reviewers of the original proposals are part of the panel to whom the presentation of completed research is made. After the presentation, the reviewers are responsible for providing a thorough critique to the program officials and the funded researchers.

From time to time, agencies mount special efforts to appraise their research programs in their entirety. For example, the Wooldridge (NIH 1965) assessment of NIH used peer review as its key procedure. Eleven panels of experts examined 240 funded external research grants and 125 unsuccessful applications, 105 training grants, and more than three dozen NIH laboratories and independent research centers. (Details of the Wooldridge study appear in Appendix C.)

Arguments For and Against Peer Judgment

Formal and informal peer judgment is the means by which science exercises continuous self-evaluation and correction. Formal peer review is the centerpiece of NSF's everyday operations and provides the agency with a key technique for performance evaluations. However, peer review is costly in terms of lost research time when it becomes formalized in such functions as proposal review, service on advisory or evaluative panels, and explicit reviews of research performance.

Peer judgment, as all human judgments, may be affected by self-interest, whatever care is taken to preclude it. Ties of friendship or association may influence judgment; so may antagonisms that have little or no bearing on the matter at hand. Some of these problems are overcome by using more than one judge. Irvine and Martin (1981, see also Martin and Irvine 1981) have suggested that, in assessing research groups --not individuals--peer evaluation be augmented by such other indicators as number of publications, citations, and highly cited papers. (These techniques are described in the section on "Bibliometric Analysis" below.) In the long run, peer judgment is corroborated by the verifiability of research findings--an external standard not available in most other areas of human judgment.

Some of the opposition to peer review is based on the perception that it relies on "insiders" who tend to favor each other's work and are resistant to new ideas.

As has been noted earlier, however, Cole et al. (1978, 1981) found the NSF review process to be operating fairly and without apparent bias--that there is in fact little evidence of an "old-boy network." Another apparent problem is the uncertainty attached to peer review in any specific instance. This is a consequence of some of the characteristics of basic research discussed in the previous chapter--for example, the eventual value and impact of a piece of research may not be apparent for some time. The 1981 Cole et al. study showed a high degree of agreement between two independent sets of funding recommendations about the top and bottom quintiles of proposals submitted to NSF. By quintile (ordered sets of 30 out of 150 proposals) and starting with the proposals rated highest by NSF reviewers, the specific rates of agreement on whether or not the proposal deserved funding were 90, 69, 56, 70, and 84 percent respectively. Reversals among proposals in the middle range are not difficult to understand since the average rating of such proposals lies near the cut-off point for funding.

Whatever the defects of peer judgment, it has worked, as evidenced by the broad record of accomplishment of the scientific establishment of which it is a central part. It will continue to be chosen by research scientists as the main process for evaluating scientific research performance. We know of nothing better. The question is in what ways, if any, it needs to be extended in the case of NSF to provide adequate evaluation of the research the agency has sponsored.

Bibliometric Analyses

Bibliometric analysis involves counts of publications and of formal citations to publications. The two types of counts reflect peer evaluations of a scientist's work, because a manuscript is published in a refereed journal only when the reviewers and editor decide that it is of sufficient merit, and because citation is recognized, at least by most scientists, as the appropriate procedure for acknowledging that the ideas, methods, or data in the cited paper influenced their own work. (For a valuable description of this process, see Zuckerman and Merton 1971.) In this context, publication counts are regarded as a measure of a scientist's productivity and citation counts as a measure of the impact of what has been produced.

Using bibliometrics to analyze the scientific literature is made practical because information on the papers published each year in most of the major scientific journals is compiled in one source document, the Science Citation Index (SCI). The information in SCI consists of the name(s) of the author(s) of each paper and their institutional affiliation(s). In addition, all works cited in each paper are listed by the name of the first author. According to Narin (1981), in 1973 the SCI covered some 5 million references contained in the more than 400,000 articles that appear in a typical year in 2,300 major scientific journals. Although the SCI was originally designed as a practical tool for conducting literature searches, it has become a means for studying the processes of science and the productivity of scientists.

Given that bibliometric measures are presumed to reflect peer judgments and that SCI makes computing such measures simple, inexpensive, and unobtrusive, evaluators and scholars of science--including Eugene Garfield, the inventor and leading proponent of the SCI (see, for example, Garfield 1979)--have been investigating whether bibliometrics could supplement or even supplant other measures of performance. For example, some 28 studies of bibliometric indicators have been reviewed by Narin (1976). Most of the studies are policy-oriented and were sponsored by federal agencies involved in basic research; the remainder were done by academic sociologists or information scientists. As an avowed advocate of bibliometric indicators, Narin asserts that the results of the studies generally support the idea that publication and citation counts can be useful to evaluators. (This study and the others summarized in this section are described in greater detail in Appendix C.)

A few studies using bibliometric measures have been funded by NSF. One example is the evaluation of its oceanography program (NSF 1980a) mentioned earlier. The conclusions reached by the NSF evaluators reflect the equivocal nature of the study's findings: "We suspect that, for broadly defined groups of adequate size, [bibliometric] ratings will not add much. They didn't in this study. However, we are by no means at the point yet of dropping ratings from post-grant evaluation." Another study looked at the productivity of a large number of chemists in American universities (DeWitt et al. 1979). It compared citation data with other indicators of

performance, such as institutional affiliation, grants, and honors. The results seem to support the claim that citations reflect other indicators of research achievement, at least to some degree.

Carter (1974) tried to determine whether bibliometric measures could serve as measures of scientific quality in evaluating NIH programs. She compared the peer rankings of a sample of proposals, initial and renewal, that had been submitted to NIH with the publication and citation records of the investigators and found only limited support for bibliometrics. The peer ratings of proposals submitted by medical or basic biological research teams correlated with several measures, including average citation counts per publication; the ratings of proposals from anatomy, surgery, or smaller clinical research teams did not.

Arguments For and Against Bibliometric Measures

Not all citations are equally significant, although bibliometric measurement treats them as such. Many citations are to routine methods or statistical designs, to modifications of techniques, or to standard data; some citations are made to caution against error. The most important citations acknowledge related work or suggest possible extensions or applications. Thus, the fact that an article receives many citations is not by itself sufficient evidence of scientific quality. Other perturbations in the number of citations are introduced by the practice of SCI to assign citation credits only to the first-named author of a publication. Also, scientists with very specialized research interests or in a discipline like anthropology, with a low rate of publication, generally receive lower ratings than colleagues in fields where frequent publication is the norm. Corrections that compensate for these and other factors have been developed by Narin and his associates (see Narin 1981); however, such adjustments increase the complexity and consequently the expense of bibliometric analysis.

Investigators familiar with the patterns of publication and citation rates generally caution against using bibliometrics to assess the performance of individual scientists or small aggregates of departmental size. For instance, DeWitt and his colleagues (1979) advise that findings based on bibliometric data should be corroborated with evidence from other sources: The "uncritical

use of citation data as a sole, or even major, criterion" yields unreasoned decisions about the allocation of resources that could affect adversely the careers of productive researchers and their laboratories. Another critic (Edge 1979) argues that a reliance on bibliometric evidence might lead to a distortion of the scientific enterprise. For example, if departments and funding agencies adopt bibliometric counts as the primary evidence of performance, scientists will find it unprofitable in terms of their professional future to do research in areas that do not offer the opportunity of immediate or frequent publication or to attend to the educational and other unmeasured contributions of research.

On the other hand, studies of funding programs or of entire research efforts that involve hundreds or even thousands of scientists have shown that bibliometric analyses generally agree with peer-based assessments (Ling and Hand 1980, Narin and Gee 1980, Riecken et al. 1981). These same studies also indicate that citation and publication patterns differ widely among disciplines and areas of research. Martin and Irvine (1981) hold that bibliometric indicators are useful only when comparable groups are being evaluated, and that judgments become trustworthy only if all measures of quality--including peer judgment and bibliometric indicators--converge.

Case Studies

Historians and sociologists of science have written case studies on scientific discoveries, in part to illuminate the research process. Scientific discoveries also have been recorded by participants and popularized by journalists and film makers. Recent case studies have focused on events in biochemistry, like the discovery of the structure of DNA (Watson 1968, Chargaff 1980), the origins of recombinant DNA research (Judson 1979), and the discovery of hypothalmic hormones (Wade 1981). There also have been studies on developments in physical chemistry (Edge and Mulkay 1974), the discovery of pulsars (Woolgar 1976), and the emergence of radio astronomy in Britain (Edge and Mulkay 1976). Industrial laboratories assemble histories and results of research and development programs they have carried out to explain the significance of their work. Corporate management tends to focus on the proportion of research efforts that

have paid off as the measure of success. Occasionally, particular research efforts that have failed are analyzed to determine whether the failure is attributable to scientific and technological shortcomings, to poor management, or to other factors, such as marketing errors, over which a laboratory has no direct control.

The case study is essentially a historical account of the social and intellectual developments that led to a key event in science. Some investigators move beyond straightforward description to attempt explanations of the importance of persons and events and how these determine subsequent developments (see Edge and Mulkay 1976 and Lemaine et al. 1976). Case studies can be used to understand the effects of institutional, organizational, and technical factors on the research process (see Mullins 1972, Collins and Harrison 1975, and Law 1976). Case studies also offer the possibility of identifying and following important outcomes of the research process that are not purely intellectual, such as the collaboration of scientists, the training of young researchers, the development of productive research centers.

Arguments For and Against Case Studies

Case studies permit the investigator to illuminate the complexity of the research process to depths that are not possible with other methods. Unfortunately, case studies are relatively lengthy and expensive. For his study of the first quarter century of molecular biology, Judson (1979) traveled throughout Europe and the U.S. for ten years, interviewing participants, observing experiments, and delving into laboratory archives and libraries. Wade's (1981) more limited investigation of the discovery of hypothalmic hormones also required several years of investigation. Indeed, it is the substantial cost of case studies in time and money that precludes their extensive use as an evaluation method.

In addition, case studies present two methodological problems. First, the validity of the results and conclusions obtained depends on the objectivity, investigative skills, and scientific knowledge of the person(s) doing the study. Accounts by participants generally highlight only the events in which they were actively involved and invariably present personalized versions. Comparing Watson's (1968) account of the discovery of the double helix structure of DNA with

Chargaff's account of events (1980) or with the biography of another participant, Rosalind Franklin (Sayre 1975), illustrates the problem. The investigator who carries out a case study must be willing and knowledgeable enough to ask key questions, to obtain accounts from informants whose perspectives are likely to differ, and to reconcile any discrepancies through personal judgment and archival evidence. Obviously, the need for thorough investigation adds to the costs of such studies.

The second methodological problem is that the findings of a case study are generally grounded in specific historical circumstances and therefore cannot be applied directly to other research settings. This makes translating the results of case studies into information that can be used by scientific planners and decision-makers problematical.

Case studies are a promising avenue for examining and eventually understanding such non-scientific aspects of the research process as social influences, institutional contexts, economic and political factors, and patterns of communications. Hence, the support of a limited number of case studies by NSF might illuminate some subtle factors that affect basic research, pursuant to the request made by the Senate Appropriations Committee. Progress is likely to be slow, though, in identifying factors that have significant impact on research and are not already recognized by research managers. On the other hand, case studies hold no promise at all as a method for routine program and project evaluation, especially in light of the time required and the costs. Moreover, it is unclear that the results of any one case study could be accepted as a guide to future federal action.

Retrospective Analysis

Retrospective studies are related to case studies in that they also try to reconstruct history. However, retrospective studies are generally concerned with multiple scientific or technological innovations rather than with only one. The goal is to identify linkages between the innovations and one particular type of antecedent event (usually either funding or prior research) rather than to interpret as many of the antecedent and contextual factors as possible, the usual aim of case studies. Retrospective analyses generally require that panels of experts or the investigators conducting the

retrospective study select significant advances in weaponry, say, or medicine or basic physics; each advance is then traced back to the events that made it possible. One motive behind early studies of this type, exemplified by Project HINDSIGHT (Sherwin and Isenson 1969) and T.R.A.C.E.S. (Illinois Institute of Technology 1968), was to justify investment in research and development in terms of its useful applications. More recent studies, like that of Comroe and Dripps (1977) for NIH, have dealt with the contribution of fundamental research to scientific or clinical advances. (For more detail on these studies, see Appendix C.)

Arguments For and Against Retrospective Analysis

Some variants of retrospective analysis may be appropriate for answering a particular type of accountability question--for example, the contribution of NSF support to a field over the long run. The method might resemble that of Kruytbosch (1978; see Appendix C) in starting with the selection of significant output by means of peer panels and then tracing back the extent of NSF funding. The output could be defined in terms of leading researchers, papers included in major review articles or monographs, articles cited in papers that receive annual awards, or significant theoretical or empirical advances in a discipline. Another type of output, represented by patents or other traceable industrial uses of basic research, is currently being looked at in an NSF retrospective study (NSF 1981). (It should be noted that the results of this study--or of any retrospective study that concentrates on commercially significant outputs--will depend on market and technological factors as well as on the quality of the basic research.) In addition to defining output, the contribution of the agency must also be defined in doing a retrospective study; it could include both direct funding through research grants and more indirect support through fellowships, purchase of equipment, institutional support, and the like. The definitions are critical, since the extent of contribution to a field measured through retrospective analysis will depend on just what output and what type of support are examined.

Because of the time needed for any type of reasonably valid historical study, including retrospective analysis, this method is not useful as a tool to provide short-term evaluations for improving research policy and management.

Prospective Analysis and Pilot Experiments

Prospective analysis involves assessment of a proposed policy or program before it is put into effect. The analysis can include estimates of anticipated costs and benefits, surveys of likely participants and other affected parties, or actual pilot tests, often including methodologically complex experimental designs (see Boruch and Wortman 1979, Boruch 1982). Not only program alternatives but also changes in procedures can be examined prospectively and tested experimentally. Such tests typically are conducted in-house and, unlike the large-scale tests of new federal programs sponsored in recent years, attract little outside attention.

NSF administrators and advisors have reported that the Chemistry Division conducts ad hoc experiments and more formal tests of proposed managerial changes. One reviewer for the Division has described his participation in an experiment that was intended to determine whether an alternative procedure--ranking of proposals by a panel of experts convened for the purpose--would yield a quality ordering of proposals different from or similar to those obtained through peer review by mail. (No significant differences appeared.) The more formal experiments include the current tests of changes in renewal procedures that link future funding more directly to the principal investigator's past performance, such as the accomplishment-based renewal procedures and the creativity extensions discussed in the preceding section on peer review (see also Appendix D). According to NSF officials, the decision whether or not to adopt these changes will be based on the opinions of administrators, researchers, and reviewers--in other words, of those whose work or research is affected by the proposed changes--rather than on any formal assessment of effectiveness.

Arguments For and Against Pilot Experiments

Pilot studies possess several advantages over evaluations conducted after a change is implemented. They are less likely to be constrained by the desires of advocates, administrators, or recipients to show that the innovation has positive or negative consequences. Also, because large-scale evaluations are difficult to

manage, the quality of results from restricted pilot investigations will be generally higher. Pilot studies offer other advantages as well. They yield more direct findings than can be obtained from trying to anticipate consequences on the basis of past experiences, permit evaluators to discover problems that could not be anticipated, and help to resolve the problems that are anticipated.

Pilot tests present some difficulties. Changes in procedurs and programs sometimes have effects that become apparent only years later. For example, any change in the level of support for recent Ph.D. recipients is likely to affect decisions about tenure that will be reached five to seven years later. Ideally, experimental designs allow for effects to be assessed over the long term before a change is made. Frequently, however, agencies cannot wait that long. The best to hope for is that agencies will continue to monitor the outcomes and be willing to reverse a decision if necessary. Another problem is the balance between the costs of conducting pilot tests and the resulting benefits. If a test is very complex, it may be both costly and ambiguous in its results. Or, the alternative being tested may be so expensive to implement that even a successful test will not make it attractive. Sometimes a new program or administrative change must (or cannot) be made for political, moral, social, legal, or scientific reasons. Under any of these conditions, not much benefit will be gained from expending the time and money necessary to conduct prospective experiments.

4 FINDINGS AND NEXT STEPS

Two questions were posed at the outset of this exploratory study:

> What method of postperformance evaluation is most useful for assessing the quality of research produced as a result of NSF support?
>
> In what ways can postperformance evaluation help improve the productivity of the research efforts supported by NSF?

Basic science supported by public funds, whether performed in universities or elsewhere, should not be exempt from the oversight that government imposes on other institutions. Neither science nor scientists can plead for special privileges on the basis that science is a unique endeavor. It is not evident, however, that any forms of postperformance evaluation that we have considered will produce better basic research. The quality of a human activity as complex, subtle, and elusive as scientific discovery is difficult to measure with any accuracy. The frequency with which research that appeared "useless" has proven valuable to science or technology constitutes a powerful argument against restricting support to mission-directed or socially oriented research. Basic science is most fruitful when it is autonomous. Research that takes place in a setting that is free and apparently unstructured is in fact governed, in Michael Polanyi's (1962) phrase, by "the republic of science."

The subcommittee believes that any additional evaluation procedures should be introduced only if they clearly enhance rather than constrict the environment in which research proceeds, and that formal techniques cannot usefully replace informed technical judgment. The principle that has governed our work and is reflected in our findings is that evaluation must serve the interest of advancing excellence in research and creating a productive research climate, while meeting the public concern for accountability.

Assessing the Quality of Research

In Chapter 2, we have pointed out the conflict between the need to obtain evaluative information that is current enough to be useful for policy and management purposes and the cumulative nature of science, which usually demands a long time period for assessing the significance of research results. Of the several evaluation methods described in Chapter 3, only peer review and bibliometrics yield information--whatever its validity--applicable over the short term to the performance of basic research. For longer term evaluations, the method of choice for establishing the quality and contribution of research is retrospective analysis. This method respects the inherent characteristics of the research process and the scientific enterprise, but the studies done to date, though instructive, have been protracted and costly. Since any additional evaluation procedure will require more time, money, and people, we consider it important to establish how much information is already at hand, and what purposes it serves. In this connection our subcommittee has reached the following conclusions:

FINDING 1: In NSF's Chemistry Division, postperformance evaluation already exists in many forms, though it is not always labeled and perceived as such.

In reviewing requests for support, and especially in reviewing requests to renew previous support, the applicant's recent accomplishments are an important part of the evaluation. Experiments by NSF with accomplishment based renewal procedures and productivity-based

grant extensions emphasize evaluation of past performance even more. NSF oversight procedures provide a means for checking these evaluative processes and should be specifically employed to ensure that reviewers explicitly consider recent research performance in making their funding recommendations, in addition to considering the general reputation of the scientist and the work proposed for the future.

FINDING 2: We believe that Congress is not effectively informed of the postperformance evaluation that NSF carries out on a routine basis. This deficiency results not from too few communications, but from the lack of summaries that focus on postperformance evaluation.

The subcommittee believes that a great deal of the information already being collected about NSF programs relates to the evaluation of research results. Apparently, the information is not effectively organized and presented in ways that meet Congressional concerns about accountability. The subcommittee recommends, therefore, that the following activities be carried out as next steps:

Activity 1: At least two other NSF research divisions should be examined to ascertain whether the experiences and practices of the Chemistry Division are typical of NSF divisions.

Of special interest is whether proposals from investigators who have had previous NSF grants represent an equally high proportion of all reviewed proposals and whether the detailed statistics kept by the Chemistry Division on division operations and support provided to the field are available for other divisions. The review should also include examination of the oversight activities and reports by the respective advisory committees. Examples of some NSF units that might be reviewed are the Behavioral and Neural Sciences Division, the Astronomy Division, or some of the divisions within the Directorate for Engineering.

Activity 2: Suitable samples of evaluative information already available should be analyzed to

explore the extent to which such material can serve the purpose of postperformance evaluation.

For example, in the course of a recent review of the Chemistry Division, its Advisory Committee collected data for each chemistry program from 35 to 50 folders, illustrating various categories of funding action. The data were used to address specific questions about the selection and assignment of reviewers, the adequacy of current levels of funding, and the distribution of funds among programs or specific areas of research. For the purpose of postperformance evaluation, it might be instructive to select folders in a similar manner and assemble and analyze the information contained in the proposals, background materials on applicants, reviewers' comments, memoranda and reports of program officers, and final reports. The main question to be addressed should be the quality of the research being supported. NSF should also examine the yearly statistics collected by the Advisory Committee and by chemistry program officials, describing the distribution of funds among individual programs and the support of chemistry by other funding agencies.

Such activities ought to make evident what additional information may be needed and how existing and new information can best be formulated to respond to concerns about the quality of NSF-supported research.

Activity 3: If the first two activities provide evidence that relevant information is available throughout NSF's research divisions, NSF should analyze why the information has not been better articulated and used.

NSF should consider how responsibilities for communication about evaluation are distributed within the agency and the reasons why NSF has not been fully successful in reporting about its performance to the satisfaction of the Congress. The analysis should lead to corrective action.

Activity 4: The use of retrospective studies to determine long-term NSF impact on a field of research should be investigated and evaluated, giving due consideration to the difficulty and cost of attaining significant results.

The subcommittee, with the help of the Committee on Chemical Sciences of the National Research Council, has assembled a list of highlights in chemistry over the last decade. What needs to be done is to find efficient ways of tracing back NSF-supported contributions to significant developments identified by the list--in, say, orbital symmetry, metal clusters, asymmetric and stereoselective syntheses, guest-host complexes, surface-modified electrodes, conducting polymers, solid-state nuclear magnetic resonance, gene synthesis, and picosecond spectroscopy. If leading researchers associated with each selected highlight can be identified, then NSF records can be used to establish whether or not, in what amounts, and at what stages the researchers were funded by NSF. The task would become considerably more difficult if information were desired as well about the share of NSF support relative to support received from other sources by the same researchers. The question to be answered is whether retrospective analysis can be made sufficiently economical to be adopted by NSF for periodic evaluation of the contribution of NSF-sponsored research to the various fields of science.

Improving Research Productivity

The same distinction between the short and long terms that we have proposed earlier in this chapter for assessing the quality of research can also be made for improving research productivity. From our exploratory investigation, we have come to some views about possibilities for improvement that can be instituted over the short term. Long-term improvement will require, as the Senate Committee has noted, a better understanding of the research process than now exists, particularly about the factors that increase or inhibit productivity.

One way of increasing research productivity over the short term is to improve the management of research support. In principle, improvement is possible at several different levels--for instance, in selecting projects, in allocating resources to subfields within a discipline and across disciplines, and in developing NSF program features that are intended to maintain a healthy science establishment by, say, assuring a steady flow of creative young researchers.

Project Selection

Much of the attention of Congress and of NSF as well has been centered on how good the decisions have been that lead to awarding or denying funds to individual researchers. This is understandable because decisions about individual proposals are the agency's primary means for advancing its mission of fostering research. Yet this process is probably the best developed and tested of all the elements that make up the doing of science. It has been honed by 30 years of experience with hundreds of thousands of proposals to NSF and other federal agencies that support scientific research and, outside that structure, in the many other contexts where peer judgment is used to decide scientific merit or promise.

As noted, there is considerable agreement among peer reviewers for some 40 percent of the proposals submitted to NSF--i.e., those that receive high ratings as well as those that receive low ratings. Proposals that fall into the middle range may be there either because reviewers agree reasonably well on mid-range scores or because reviewers disagree and assign either high or low scores, resulting in a mid-range average. Proposals in the middle range that fall into the second category deserve special consideration, because they may represent the very type of risky effort that should be supported with public funds.

Since the cut-off point for NSF funding of proposals ranges around 4 (out of a score of 1 to 5, with 5 being excellent, 4 very good, etc.), and researchers who submit proposals to NSF are already a self-selected group willing to compete, practically all the research that is considered for funding at any time is likely to be "very good" or better, and decisions at the margin are difficult to make. From our observation of the operations of the Chemistry Division, we find that the latitude given to the NSF staff in this process is exercised with great care. Staff recommendations (reviewed at several levels higher up) as to which "very good" proposals to fund when there is not sufficient money for all of them are based on such criteria as the state of the field pertaining to the research, whether other agencies support that type of research, how the proposed work relates to research already being funded by NSF, and so forth. Such considerations are not always

documented as well as they might be in the sense of making accessible to an outside observer the reasoning that goes into a decision. To some degree, this is also true of some reviewer responses--a possible reason for the GAO (1981) finding that performance on the preceding NSF grant was not explicitly taken into account in reviews of renewal proposals. Nevertheless, judging from our understanding of the proposal evaluation process in the Chemistry Division, we consider it to be competently and carefully managed.

FINDING 3: The subcommittee considers peer review to be the best way of choosing among individual research proposals, even though experts may occasionally disagree about the merits of specific proposals. No additional methods of postperformance evaluation that we know of will significantly improve the selection of individual projects.

The Chemistry Division has been experimenting with alternative forms of proposal review, especially for renewal proposals. We commend such experiments because they demonstrate good management. But in stating this, we do not mean that the experiments could not be improved. Improvements could be made by using stronger standards of evidence and making clear what standards are being used in assessing the effects of the experiments. Though it may require some staff effort, the information assembled may well prove useful both for managing research support and for increasing knowledge about the process. As has been noted, both the chemistry Advisory Committee and program officers assemble a great amount of objective data about the performance of the Chemistry Division. Care should be taken that applicable data are used to benefit the design and evaluation of future experiments and that findings are communicated to parties concerned with the operations of NSF.

The suggested study of other divisions should establish whether the operations there are characterized by the inventiveness and flexibility seen in the Chemistry Division. We recommend an additional step:

Activity 5: NSF should continue its small-scale experiments with management improvement. It

should assess the level of documentation necessary to evaluate and communicate the results of the experiments to interested groups.

Allocation of Resources Among Subfields

Looking at the support of chemistry within NSF, we believe that a question of great importance concerns the basis for the allocations made to subfields or programs within a division. This question, in the subcommittee's judgment, is potentially of much greater significance to the management of NSF than modifications to project selection.

FINDING 4: Additional postperformance evaluation of research should be used where the highest leverage for improvement of the agency's performance can be obtained. The quality of NSF-supported research should be assessed in some aggregate form--for instance, in analyzing the allocation of resources among the subfields of a discipline or between disciplines.

Individual programs within the Chemistry Division change slowly. Allocations among programs vary little from year to year. However, the type of research undertaken within a program may shift considerably over time. Because grants in chemistry are relatively modest ($60,000 on the average), it is possible to fund some exploratory lines of research without enormous risk, and--in case of success--to increase subsequent funding to the point where a new subspecialty has become part of a program. From time to time, the Chemistry Division has organized workshops to identify research needs and opportunities in subfields such as crystallography, high-temperature chemistry, lasers, and physical organic chemistry in an effort to attract chemists to a particular area of research. Evaluating the success of such efforts might be one way of determining whether this is a useful means for encouraging work in promising areas of research.

How are decisions made to increase or decrease funding in a subfield? To what extent is the relevant research community involved? In order to appraise the

need for change, the output and quality of programs must be assessed in the aggregate. The appropriate method is some form of peer judgment. Thus,

Activity 6: Various alternatives should be explored for using outside experts to appraise the aggregate quality and results of research supported by NSF in each program within a research division.

The appraisal should include advice on needed changes in direction. Possibilities include visiting committees, working conferences, or less costly versions of the Wooldridge assessment of the National Institutes of Health (NIH 1965). Whatever the procedure, it must be such that the advice goes to the Director and the assistant directors of NSF, as well as the division directors. In principle, similar reviews could serve to improve allocations among fields, but we recognize that the difficulties of making such decisions increase as comparisons between substantive areas of research become more difficult.

Maintaining Scientific Strength--Young Researchers

NSF was charged at its inception in 1950 with the goal of promoting "the progress of science." In subsequent years this objective has come to include responsibility for maintaining and improving the institutional and organizational capacity of the U.S. research community and for meeting the nation's requirements for scientific personnel. Some important and complex issues are subsumed under the former--for example, the balance of funding for instrumentation, technical support staff and research scientists, and the degree of concentration of institutional resources at the leading research universities. The subcommittee discussed several of these issues but had time to concentrate on only one: NSF's responsiblity for meeting the nation's requirements for scientific personnel--specifically, the development of a continuing supply of able young researchers.

The availability of talented and well-trained young researchers is fundamental to the health of the scientific community. New scientists in such fields as

mathematics and physics are known to be responsible for many of the innovations in theory and techniques (Kuhn 1970, Cole et al. 1973). Equally important, their presence in universities and in industrial research and development units is an essential factor in ensuring that the quality of science will not decline in the decades to come. NSF can and does contribute to the support of young scientists at three stages: in providing opportunities for their formal education and training as graduate students and postdoctoral fellows, in funding initial proposals for independent research, and in supporting requests for grant renewals. As to the support of initial research proposals, data collected by the Chemistry Division (NSF 1980b) for 1973 through 1979 show that young investigators are somewhat more likely than established investigators to receive funding for a proposal for new research. Similarly, Cole et al. (1978) report that young researchers are not at a disadvantage in getting NSF grants: For 1,200 proposals drawn from economics, solid-state physics, and chemical dynamics, "[a]ge had no strong effect on either ratings received or the probability of receiving a grant [authors' emphasis]."

Nevertheless, some questions remain about NSF support for young researchers. The data from the Chemistry Division are not displayed so that we can readily trace decisions on renewal proposals submitted by researchers who were "young" (seven or fewer years after their doctorate) when they got their first award but no longer fall into that category after completing their original three-year grant. Therefore, it is not clear to the subcommittee what provisions NSF makes during the renewal proposal stage for young investigators.

Activity 7: Additional information should be assembled about the support of young researchers, particularly at the first renewal stage.

The objective of supporting young researchers is to ensure their entry into the field and the transition of the best to the status of established and productive scientists. It is important to find out how policies and procedures within the chemistry programs and other NSF programs are affected by the findings about young researchers and by the projections about the future requirements for research scientists in U.S. universities

and industries. Periodic checks of NSF performance will be needed so as to ensure that the agency's programs and procedures meet changing national needs.

Long-term Improvement

The Congressional request stressing the need to identify factors that make for research success recognizes the complexity of the research process and the difficulty of improving predictions about it. Intensive case studies and a variety of surveys of scientists have been carried out to develop greater understanding of the effects of institutional, organizational, and technical factors on the research process. While such work has been useful in illuminating specific discoveries, it has not advanced knowledge to the point of identifying generally applicable productivity factors beyond those already used in evaluating research. The identification of additional factors will require protracted and extensive research which may, in the end, yield little of practical utility. The question as to how much more understanding we can gain about the research process and how useful such added knowledge will be in improving NSF's funding decisions can be pursued through support of a few additional, judiciously chosen case studies of the context and events that have accompanied specific scientific advances.

REFERENCES

Boruch, R.F. (1982) Experimental Tests in Education: Recommendations from the Holtzman Report. American Statistician 36 (February): 1-8.

Boruch, R.F., and Wortman, P.M. (1979) Implications of Educational Evaluation for Evaluation Policy. In David C. Berliner, ed., Review of Research in Education. Washington, DC: American Educational Research Association.

Bronowski, J. (1956) Science and Human Values. New York: Julian Messner, Inc.

Bush, V. (1945) Science--The Endless Frontier: A Report to the President on a Program for Postwar Scientific Research. Washington, DC Office of Scientific Research and Development.

Carter, G.M. (1974) Peer Review, Citations, and Biomedical Research Policy: NIH Grants to Medical School Faculty. R-1583-HEW. Santa Monica, CA: Rand Corporation.

Chargaff, E. (1980) Heraclitean Fire: Sketches from a Life Before Nature. New York: Warner Books.

Chemical and Engineering News (1980) NSF Chemistry Funding Draws Praise. (August 18): 19-20.

Clark, R.W. (1961) The Birth of the Bomb. London: Phoenix House, Ltd.

Cole, J.R., and Cole, S. (1973) Social Stratification in Science. Chicago: University of Chicago Press.

Cole, J.R., and Cole, S., with the Committee on Science and Public Policy (1981) Peer Review in the National Science Foundation: Phase II of a Study. Washington, DC: National Academy Press.

Cole, S., Rubin, L., and Cole, J.R. (1978) Peer Review in the National Science Foundation: Phase I of a Study. Washington, DC: National Academy of Sciences.

Collins, H.M., and Harrison, R.G. (1975) Building a TEA Laser: The Caprices of Communication. Social Studies of Science 5 (4): 441-450.

Comroe, J.H., Jr. (1977) Retrospectroscope. Menlo Park, CA: Von Gehr Press.

Comroe, J.H., Jr., and Dripps, R.D. (1977) The Ten Top Clinical Advances in Cardiovascular-Pulmonary Medicine, 1945-1975. NIH 78-1522. Two Volumes. Washington, DC: Department of Health, Education & Welfare.

Crane, D. (1972) Invisible Colleges: Diffusion of Knowledge in Scientific Communities. Chicago: University of Chicago Press.

DeWitt, T.W., Nicholson, R.S., and Wilson, M.K. (1979) Science Citation Index and Chemistry. Scientometrics 2 (4): 265-275.

Edge, D.O. (1979) Quantitative Measures of Communications in Science: A Critical Review. History of Science 17: 102-134.

Edge, D.O., and Mulkay, M.J. (1976) Astronomy Transformed: The Emergence of Radio Astronomy in Britain. New York: John Wiley and Sons.

Edge, D.O., and Mulkay, M.J. (1974) Case Studies of Scientific Specialties. Working paper. Science Studies Unit, University of Edinburgh, Scotland.

Franks, F. (1981) Polywater. Cambridge, MA: MIT Press.

Garfield, E. (1979) Citation Indexing. New York: John Wiley and Sons.

Gaurman, J. (1980) Rating of Graduate and Professional Programs in American and International Universities. National Education Standards.

General Accounting Office (1981) Better Accountability Procedures Needed in NSF and NIH Research Grant Systems. PAD-81-29. Washington, DC: U.S. General Accounting Office.

Holton, G. (1978) The Scientific Imagination. Cambridge, England: Cambridge University Press.

Illinois Institute of Technology (1968) Technology in Retrospect and Critical Events in Science. NSF-C535. Chicago: Illinois Institute of Technology Research Institute.

Irvine, J., and Martin, B.R. (1981) Internal Criteria for Scientific Choice: An Evaluation of the Research Performance of Electron High-Energy Physics Accelerators. Science Policy Research Unit, University of Sussex, England.

Judson, H.F. (1979) The Eighth Day of Creation. New York: Simon and Schuester.

Kruskal, W. (1975) Draft Statement about Prospective Studies. University of Chicago.

Kruytbosch, C. with Papenfuss, S. (1978) Some Social and Organizational Characteristics of Breakthrough Science: An Analysis of Major Innovations in Four Fields of Science, 1950-1976. Paper presented at the IXth World Congress of Sociology, Uppsala, Sweden. National Science Foundation, Planning and Policy Analysis Office.

Kuhn, T.R. (1970) The Structure of Scientific Revolutions. Revision of 1962 Edition. Chicago: University of Chicago Press.

Law, J. (1976) The Development of Specialities in Science: The Case of X-ray Protein Crystallography. In Lemaine, G., Macleod, R, Mulkay, M., and Weingart, P., eds., Perspectives on the Emergence of Scientific Disciplines. Chicago: Aldine Publishing Company.

Lemaine, G., Macleod, R. Mulkay, M., and Weingart, P., eds. (1976) Perspectives on the Emergence of Scientific Disciplines. Chicago: Aldine Publishing Company.

Ling, J.G., and Hand, M.A. (1980) Federal Funding in Materials Research. Science 209 (September 12): 1203-1207.

Mac Lane, S. (1980) Total Reporting for Scientific Work. Science 210 (October 10): 158-163.

Martin, B.R., and Irvine, J. (1981) Assessing Basic Research: British Optical Astronomy and the Isaac Newton Telescope. Science Policy Research Unit, University of Sussex, England.

Morrison, P. (1957) The Overthrow of Parity. Scientific American 196 (April): 45-53.

Mullins, N.C. (1972) The Development of a Scientific Specialty: The Phage Group and the Origins of Molecular Biology. Minerva 10 (January): 51-82.

Narin, F. (1981) Concordance Between Subjective and Bibliometric Indicators of the Nature and Quality of Performed Biomedical Research. Unpublished Ph.D. dissertation. Walden University, Naples, FL.

Narin, F. (1976) Evaluative Bibliometrics: The Use of Citation Analysis in the Evaluation of Scientific Activity. Cherry Hill, NJ: Computer Horizons, Inc.

Narin, F., and Gee, H.H. (1980) An Analysis of Research Publications Supported by NIH, 1970-1976. Washington, DC: Public Health Service, National Institutes of Health.

National Commission on Research (1980) Accountability: Restoring the Quality of the Partnership. Washington, DC: National Commission on Research.

National Institutes of Health (1965) Biomedical Science and Its Administration. (Wooldridge Report.) Study Prepared for the White House. Washington, DC: National Institutes of Health.

National Science Board (1981) Science Indicators 1980. NSB 81-1. Washington, DC: U.S. Government Printing Office.

National Science Foundation (1981) Impact on Industrial Innovations Associated with Chemistry. Draft of a Report. Washington, DC: National Science Foundation.

National Science Foundation (1980a) Evaluation Study of NSF's Oceanography Program. Washington, DC: National Science Foundation.

National Science Foundation (1980b) Oversight Review Report for Chemistry. Final Report. Washington, DC: National Science Foundation.

National Science Foundation (1979) Administration and Management. Subject: External Peer Oversight. Circular No. 147. Washington, DC: National Science Foundation.

Plott, C.R. (1974) A Review of Decision Theoretic Literature with Implications Regarding Governmental Research and Development Policies. In Government Policies and Technological Innovation. Vol. 2: State-of-the-Art Surveys. PB-244572/AS. Washington, DC: National Technical Information Service.

Polanyi, M. (1962) The Republic of Science: Its Political and Economic Theory. Minerva I (Autumn): 54-73.

Popper, K.R. (1959) The Logic of Scientific Discovery. New York: Basic Books.

Price, D.K. (1978) Endless Frontier or Bureaucratic Morass? Daedalus (Spring): 75-92.

Riecken, H.W., Feldman, J.S., and Zelinger, G. (1981) Citation Analyses for Evaluating A Biomedical Research Program: A Bibliometric Study of VA and NIH Research Output. School of Medicine, University of Pennsylvania.

Sayre, A. (1975) Rosalind Franklin and DNA. New York: Norton.

Sherwin, C.W., and Isenson, R.S. (1969) Final Report on Project HINDSIGHT. Washington, DC: Office of the Director of Defense Research and Engineering, Department of Defense.

Staats, E. (1979) Federal Research Grants. Science 205 (July 6): 18-20.

Streitwieser, A., Jr. (1981) The 1981 Nobel Prize in Chemistry. Science 214 (November 6): 4521.

Thomas, L. (1974) The Lives of a Cell. New York: Viking.

U.S. Senate (1980a) Department of Housing and Urban Development - Independent Agencies Appropriations Hearings. 96th Congress, 2nd Session. Washington, DC: U.S. Government Printing Office.

U.S. Senate (1980b) Department of Housing and Urban Development - Independent Agencies Appropriation Bill, 1981. 96th Congress, 2nd Session. Report No. 96-926. Washington, DC: U.S. Government Printing Office.

U.S. Senate (1979) Department of Housing and Urban Development - Independent Agencies Appropriation Bill, 1980. 96th Congress, 1st Session. Report No. 96-258. Washington, DC: U.S. Government Printing Office.

U.S. Senate (1978) Department of Housing and Urban Development - Independent Agencies Appropriations Hearings. 95th Congress, 2nd Session. Washington, DC: U.S. Government Printing Office.

Wade, N. (1981) The Nobel Duel. Garden City, NY: Anchor Press/Doubleday.

Watson, J.D. (1968) The Double Helix. New York: Atheneum.

Wilson, L.S. (1980) Government-University Relationships: The Conduct of R&D. Washington, DC: National Technical Information Service.

Woolgar, S.W. (1976) Writing on Intellectual History of Scientific Development: The Use of Discovery Accounts. Social Studies of Science 6: 395-422

Yang, C.N. (1961) Elementary Particles: A Short History of Some Discoveries in Atomic Physics. Princeton, NJ: Princeton University Press.

Ziman, J. (1968) Public Knowledge: The Social Dimensions of Science. Cambridge, England: Cambridge University Press.

Zuckerman, H., and Merton, R.K. (1971) Patterns of Evaluation in Science: Institutionalization, Structure, and Functions of the Referee System. Minerva 9 (1): 66-100.

APPENDIX A

EVALUATION OF BASIC RESEARCH IN INDUSTRY

U.S. industry supports basic research in the physical sciences in general--and chemistry in particular--in order to stimulate new commercial products and more efficient manufacturing processes. Of the $8.8 billion spent on basic research in the United States during 1981, industry spent about $1.5 billion--double the outlay of universities and colleges, and about one-fifth as much as the federal government with its large expenditures for defense and space research. Yet basic scientific research constitutes for most corporations only a small part of the total research and development effort. For example, the research division of IBM employs some 2,000 professional or technical personnel, only one-quarter to one-third of whom deal with problems comparable to those of academic scientists. Annual expenditures for basic scientific research at Merck, one of the nation's largest pharmaceutical houses, amount "to only a few percent" of the research and development budget. With relatively few resources given over by corporations to fundamental science, the procedures used by the managers of industry research divisions to evaluate the progress and results of basic research become an important means of ensuring that the best possible use is made of what resources are available.

Accordingly, the subcommittee contacted managers of research divisions at several major corporations and asked them to describe their procedures for assessing basic research and the rationale for doing this. One of the managers, A.M. Clogston of Bell Telephone

Laboratories, Inc., provided the following paper, which describes in some detail the problems of assessing scientific performance in an industrial setting and how Bell's management has dealt with those problems.

* * *

POSTPERFORMANCE EVALUATION OF RESEARCH AT BELL LABORATORIES

The need for retrospective evaluation of research has been recognized for a long while at Bell Laboratories, and well established procedures have been developed to carry out such evaluations. They have stood the tests of time and usefulness in managing research programs and evaluating individual research performance. No analogous procedures seem to have ever been applied systematically to postperformance evaluation of research at universities or other institutions working under contract with federal granting agencies, but the pressure for some degree of postperformance evaluation is evidently growing in Congress and becoming of increasing concern to the agencies. The Bell Labs procedures and context for research evaluation are described below as a contribution to a study of possible methodology for postperformance evaluation of government-funded research. However, it is not at all apparent that the experience of Bell Labs is easily transferable to research funded and evaluated by an agency external to the institution in which the research is carried out.

Evaluation of research implies some criteria against which the research is to be evaluated. These are by no means as obvious as they might seem at first, because research is not carried out as a series of prescribed tasks whose successful completion can be measured. Since the product of research is knowledge, and it is the quality and applicability of knowledge that is being evaluated, the criteria are necessarily more complex. They are generally of two kinds--measuring either the impact of research on the advancement of science or measuring its impact on the discovery and applications of new technology. Moreover, there also are short- and long-range aspects that must be taken into account. In view of the complexities, the procedures

for making sound evaluations of research need to be reasonably sophisticated and disciplined. Understandably, the procedures present difficulties for the development of postperformance evaluation procedures by government bodies that make research grants.

What I describe below about research evaluation at Bell Labs therefore has two aspects--the criteria for evaluation and the procedures by which evaluations are carried out. I will start with a discussion of procedures, not only because this is probably the most distinctive aspect of the Bell Labs process, but also because it will establish the context for the discussion of criteria.

Research at Bell Laboratories is evaluated by line management at several different levels and in several different ways. The most formal process is an annual review of individual performance. Almost always this is first carried out at the laboratory level by the laboratory director and his five or six department heads. (A laboratory under a Research Vice-President at Bell Labs is typically an organized unit of about 50 research scientists assisted by 40 or so research associates.) The director and his department heads spend several days each year carefully reviewing the past year's work of each scientist and evaluating it in terms of the degree to which it meets expectations based on the evaluative criteria to be discussed below. The essential feature of these evaluations is that the director and heads, who have strong personal research credentials, are intimately informed about the work of each scientist. Based on the experience at Bell Labs, such a well informed management group is able to arrive readily at a consensus view that has general acceptance. In order to establish overall consistency in the research area, the laboratory-level performance evaluation is followed by less detailed reviews conducted for each individual at division and vice-presidential levels. The results of the reviews are later incorporated into pay increases in such a way that salary reflects performance.

The annual performance review is supplemented by periodic informal reviews of ongoing research presented by each scientist to his management through the division level. These reviews are a principal means by which management can be aware of the significance and progress of research work, thereby contributing to the soundness of

the evaluative process. Other elements involved in performance appraisal are indicators of outside recognition accorded a scientist, such as invited papers at conferences, prizes, and the general level of outside collaborations. Still another element is the quality of the contribution made to the scientific literature, though individual publications are not usually evaluated as such, nor is undue weight given to the number of publications. No use is made of citation analysis.

There are two features of this evaluation process that should be pointed out. First, the evaluations are based on ongoing research and not necessarily on completed pieces of work. Research, in fact, is usually an ongoing enterprise, so that the idea of research projects with a definite beginning and end is somewhat artificial. Although research often has an episodic character punctuated by the publication of papers, published papers alone have never served as an adequate basis for research evaluation at Bell Labs and would not fit well with its concept of an annual review of performance. Secondly, research evaluation at Bell Labs is a matter of the total impact of a scientist on science and technology considered important to the company and not just an evaluation of the published research record. Other considerations that are important are the scientist's leadership and influence on others, his ability to couple new science into Bell Labs, his ability to orchestrate and integrate research requiring contributions from several individuals, and his general level of productivity, orginality and capability for independent work. The Bell Labs evaluation process requires that the assessments be carried out internally by people well acquainted with each scientist's research. This essential feature of the process would not seem to be easily transferable to a context in which research evaluation is done by an external agency. It also seems apparent that postperformance evaluation of research carried out only on the basis of published work will miss important elements of a scientist's contributions whether the researcher is in industry or at a university.

I turn now to the second aspect of the evaluation process--namely the criteria against which evaluations are made. In the broadest sense, research at Bell Laboratories is evaluated in terms of its impact on our mission to supply the technology the Bell System needs to do its job in both the short and long term. Bell

Labs' research, therefore, can have value both for its short-term impact on communications technology and for its long-term impact through the developing body of knowledge we expect to be important for future Bell System technology.

The short-term impacts of research are relatively easy to identify and evaluate. They will generally relate to well-developed technology known in our case to be important for communications; they will be important of and by themselves without need for other supporting developments; and there will generally exist some route by which the research results can be incorporated into operating technology within a reasonable length of time. In applying the criterion of short-term impact, two conditions are obviously necessary. First, the important technologies need to have been identified and accepted by the organization, and second, the group that evaluates the research must be fully aware of the technologies.

The long-term consequences of research are harder to identify and evaluate because they require judgments about such matters as originality, creativity, significance, sophistication, reliability, and relevance to long-range goals. The most common view of evaluating research is probably one of assessing the effects of a piece of research on its own narrow field of specialization--or perhaps on the somewhat broader area of science in which it is embeded. This is really a judgment about short-term impact and therefore is relatively easy to make. But it leaves unanswered, for example, the expected long-term impact of a specialized field of science on the goals of the organiation, either by itself or through its effects on a broader area of science. These can be difficult matters to decide in a context where research may have to be evaluated without a full appreciation and definition of the technology it is expected to serve. In the case of an agency like the National Science Foundation it may amount to no more than a "feeling" that high-energy physics, for example, is an important area in which the United States should conduct research. The difficulty of making long-range evaluations in such cases is often expressed as the need to await the judgment of history.

Evaluating research for its long-range impact is made easier at Bell Laboratories because a conscious effort is made to identify the fields of science and specializations that are considered to be important for future Bell Labs technologies. A research evaluation

that points to an important impact on one of these fields of science can therefore be presumed to have a valuable long-term impact on Bell Labs' mission.

Other measures of potential long-range impact, such as originality, creativity, significance and sophistication, can generally be evaluated on the internal evidence of a completed piece of research. They form an important part of the Bell Labs evaluative process and are highly valued because research performed with those qualities is very much more likely to have a lasting, long-term impact. Concerning the question of reliability of a piece of research and its ability to stand the test of time, there is no substitute for close familiarity of the evaluators with the research being assessed.

It should be evident that the procedures described here are closely tailored to the Bell Labs situation and may not be easily transferable to other industrial or government settings. They may, however, contribute to the general understanding of how research can be effectively evaluated in one industrial context and thereby provide lessons for use in other circumstances.

A.M. Clogston

* * *

Evaluation at Other Firms

The managers of research divisions at other firms who responded to the request of the subcomittee for information made several additional points about the evaluation of basic research in industry, which can be summarized as follows:

On the whole, industry conducts its evaluation of basic research at three levels: the individual scientist, aggregates of several to several dozen scientists (sometimes referred to as laboratories, projects, or research units) who work together on the same or related problems, and the entire division or department of scientists who are engaged in basic research. The emphasis on evaluation at each level varies widely among firms, and even firms that stress the same levels of evaluation rely on different procedures and practices.

For example, DuPont, a firm which, like Bell Laboratories, assesses the performance of individual employees, does so with criteria and measures different from those used at Bell Laboratories. The Bell evaluation highlights the ability of employees to facilitate research, as well as their scientific accomplishments, while evaluation at DuPont focuses on contributions to scientific knowledge. Evidence on DuPont scientists is assembled from periodic oral and written reports provided by the scientists themselves and from external sources--patents and publications, citation counts, invitations to lecture or write reviews, and comments from scientists at universities who work in the same field. The evidence is assessed annually by supervisors and managers.

Whatever the criteria and measures used to assess individuals, the judgments that result become a major factor in personnel decisions made at several of the firms whose managers were contacted. Three major elements of the professional careers of DuPont scientists --their compensation, chance for promotion, and freedom to pursue research problems of their own choosing--depend primarily on the results of the annual evaluation. Similarly at Merck, where the middle management annually reviews the fundamental research that the firm supports, the scientists in charge of the research are rewarded "if scientific progress is being made."

The firms that evaluate the performance of laboratories or research units typically do so, according to their managers, for one or both of the following reasons: (1) to mark laboratories or units that, on the basis of projected productivity, should have their support either enlarged or reduced and (2) to identify results that have progressed far enough for them to be taken over by other units that specialize in the development of new products or manufacturing technologies. Once a research unit is established, considerable time and effort may be spent monitoring the course of the research effort and its results. Such assessment is practical in the case of industry because managers--who began their professional careers as scientists and then acquired administrative responsibilities--maintain regular contact with the units under their supervision. In addition, some firms require oral or written reports annually from all units as well.

Decisions about continuing support for laboratories or research units are based on judgments about the level

of progress that the unit has achieved. A verdict to discontinue support is reached, however, only after sufficient time has passed to ensure that no practical results will be forthcoming. At Merck, for instance, "five to ten years--or even longer--would be the average duration of a fundamental research project before the whistle is blown." In the meantime, as applications of research results are perceived, managers spend time and effort in assessing the projected costs and benefits before recommending commercial development.

Considerable time and effort also is spent by the management at some firms--and by the management at Exxon in particular--in deciding whether to move into new fields of basic research. Factors that enter into the prospective, or "front end," assessment of such fields at Exxon include the state of the science and technology that are currently employed to meet the needs of the firm, the level of progress that has already been achieved in the fields proposed for expansion, and the knowledge and skills required for an additional research unit to be productive. In other words, according to A. Schriesheim, "We [at Exxon] do not set up long-range research groups lightly, and we do not abandon them precipitously."

Only two of the research managers we contacted report that their firms evaluate the units engaged in fundamental research as a whole. The assessment of the Central Research and Development Department of DuPont, which is done retrospectively, takes into account the long periods of time that may pass from the point of an initial discovery to its commercial application. Thus, managers are requested periodically to identify and tabulate the new products and improved manufacturing processes resulting from basic research at DuPont over a ten-year period. By contrast, the retrospective evaluation of IBM's research division focuses on the short term. Directors of the research units report annually to top management on the most significant results that came to light during the previous year.

The emphasis of the division- or department-level evaluation is on identifying "winners"--that is, basic research supported by the firm with some real or anticipated commercial significance. Top executives at IBM and DuPont do not expect that every project they support will be successful, because they recognize the uncertainties of basic research. "One starts many things [in industrial laboratories that support basic research]," according to

Ralph Gomory of IBM, "in order to have some successes." What the retrospective analysis does offer is evidence that the basic research units are achieving an acceptable number of successes. Since the accomplishments in scientific research at both firms in recent years have been described by executives as "impressive," neither IBM nor DuPont has deemed it necessary to review its basic research activities more extensively.

In summing up, Howard Simmons of DuPont observes that "we [the managers of fundamental research in industry] recognize the high risk and low yield of fundamental research, we place our bets on outstanding individuals, and we recognize that the significance of fundamental research may not become illuminated until many years after the work has been accomplished."

* * *

The subcommittee wishes to thank the following individuals who responded in writing to our request for information about the evaluation procedures used in industry: A.M. Clogston, Executive Director, Research, Physics and Academic Affairs Division, Bell Laboratories; Ralph E. Gomory, Vice President & Director of Research, IBM International Business Machines Corporation; Lewis H. Sarett, Senior Vice President, Science and Technology, Merck & Company, Inc.; A. Schriesheim, General Manager, Engineering Technology Department, EXXON Research and Engineering Company; Howard E. Simmons, Jr., Director, Central Research & Development Department, E.I. du Pont de Nemours & Company, Inc.; and L.J. Thomas, Director, Research Laboratories, Eastman Kodak Company.

APPENDIX B

HOW SCIENCE JOURNALS PICK PAPERS: SUMMARY OF A SURVEY

Lawrence S. Wolfarth

The editors of almost all major scientific journals base their decisions as to what papers to publish using peer-based judgments of scientific merit. Peer review procedures date from the publication of the first scientific journals in the seventeenth century (Zuckerman and Merton 1971), and modern scientists have come to regard the process as critical to ensuring a high quality of scientific research. It has been characterized by the English physicist, John Ziman (1968):

> The fact is that the publication of scientific papers is by no means unconstrained. An article in a reputable journal does not merely represent the opinions of its author; it bears the imprimatur of scientific authenticity, as given to it by the editor and the referees he may have consulted. The referee is the lynchpin about which the whole business of Science is pivoted.

Given the charge of assessing methods for postperformance evaluation, the subcomittee wanted to understand better the structure and operation of the review system used by scientific journals for evaluating the knowledge gained from basic research. Accordingly, editors of the principal journals sponsored by the American Chemical Society (ACS) were contacted and asked to describe how manuscripts from scientists are evaluated

and decisions to publish are made. Eleven editors of ACS-affiliated journals and the editor of the nonaffiliated Journal of Chemical Physics replied. Their responses are summarized in this report, which covers such matters as acceptance rates, the process of selecting reviewers, guidelines for reviews and their application, and the process of making decisions to publish or not.

Acceptance Rates for Manuscripts

The statistics provided by the editors about the fate of papers submitted to their respective journals indicate that chemical journals are moderately selective. During 1980 (or, for some of the journals, a period of several years ending in 1980), the rate at which papers were rejected outright ranged from 20 percent to 52 percent, with half of the journals rejecting 30 percent or more. These figures are supported by data from the annual report of the American Chemical Society (Chemistry and Engineering News, April 13, 1981), which indicate that two manuscripts are published for every three received.

Nearly all manuscripts accepted by the chemical journals on which we had information required some revision before they were deemed publishable. The likelihood of a paper appearing without need for any revisions ranged from zero to 10 percent. Perhaps not surprisingly, several editors expressed dismay when reporting the percentages; they indicated that the problem of revisions reflects an inability of many scientists to communicate ideas clearly and logically. Nonetheless, the changes required for most manuscripts were minor. One editor reported that 58 percent of all manuscripts sent to his journal were accepted after a single set of modifications; two other editors observed that only 25 to 30 percent required a second revision.

The Selection of Reviewers

Implicit in the accounts of how editors assign reviewers was a primary concern that reviewers need to be knowledgeable enough about the topics and methods that are covered in the paper to be able to gauge its significance and to make any appropriate technical criticisms.

For this reason, reviewers are generally found through professional interactions that allow the editor to judge firsthand the knowledge and abilities of prospective reviewers. For instance, one editor claimed that he first met most of his reviewers at scientific conferences. Two of the editors stated that they rely on scientists who recently published in their journal, arguing that scientists who have been exposed directly to editorial standards and preferences will probably produce better reviews. Young researchers are employed as reviewers by another editor whenever possible because "very often the leading scientists in an area are so overwhelmed by demands on their time that they cannot read papers carefully."

For particular papers, depending on the editor, the list of potential reviewers may be supplemented either by including scientists whose work is cited in the paper or by requesting authors to submit the names of scientists who are knowledgeable about the area of research. In such cases, editors--for that matter, most scientists--presume that personal ties cause no problem of bias or favoritism because they assume that reviewers will provide an even-handed assessment of the paper as a matter of course. (For evidence supporting this assumption, see Zuckerman and Merton 1971.) Should the rating or comments suggest that a reviewer is biased or plainly prejudiced, an editor is likely to remove the offending scientist from the list of prospective reviewers, according to those editors who took part in the survey.

The assignment of specific reviewers to a manuscript, for nine of the ten journals on which we had information, is done informally. After a manuscript is received at the journal's office, an editor reviews the abstract so as to decide whether the topics covered are pertinent to the readers. If they are, the editor then identifies two or three scientists from the list of potential reviewers whose professional interests and research match those of the paper. The scientists are asked to take on the assignment if they can meet the deadlines.

The single journal that has formalized its system of assigning reviewers is Biochemistry. The basis of the system is a questionnaire sent in 1979 to some 3,500 life scientists, asking them to classify their research interests from among 75 categories. (See Garson 1980 for details.) The responses were transferred to a computer

file that now contains the name, address, and five principal interests of each respondent. When a manuscript is received by Biochemistry, it is coded as to subject content and research methods, and this information is entered into a computer. The computer then compares the codes to the stored information and produces the names of two suitable reviewers, along with their current addresses and records of previous reviews. It is not yet clear whether the computerized system yields better reviews or distributes the burden of reviewing more equitably or both, but computerization has facilitated one important task in running a journal.

Among the problems faced by editors in dealing with reviewers is finding scientists who are conscientious about standards and deadlines and capable of evaluating any paper regardless of its complexity or scope. The best reviewers are generally distinguished researchers whose schedules are apt to be filled with professional commitments. Two of our respondents state that they deliberately limit to four or less the number of manuscripts sent to top scientists in a given year so as to avoid overburdening them. Another more serious problem is the reviewer who fails to meet deadlines (ranging from two weeks to several months) or who returns vague or incomplete comments that suggest he never read the paper carefully. The offending reviewers are usually eliminated from the list of potential reviewers when such misbehavior persists.

Guidelines for Reviewers

The ratings of manuscripts by even the most conscientious reviewers are likely to diverge because of the variety of criteria on which papers can be judged. Editors can help to ensure that the focus of the reviews is comparable by setting out the criteria that ought to guide the evaluation of papers and by communicating these criteria in writing to reviewers at the time the manuscript is sent. We asked editors to provide us with their personal views about the guidelines for reviewing. We also requested copies of any formal response form used by reviewers as well as any guidelines provided to authors and reviewers, giving the technical, substantive, and other requirements for manuscripts and the criteria for publication.

Editors stressed three criteria in answering our question about guidelines: scope (mentioned specifically by five editors), adequacy of the experimental work described (mentioned by five), and the significance of the research (also mentioned five times). The first concerns whether or not a manuscript is likely to appeal to readers of the journal. Thus, an editor for a nonspecialized publication like the Journal of the American Chemical Society (JACS) looks for papers that cut across traditional divisions between fields in chemistry. An editor for specialized journals such as Biochemistry or Organometallics prefers papers that deal with theoretical or methodological issues that are specific to a field. Articles that describe results or techniques with obvious commercial significance are preferred by editors of applied chemistry journals. The second criterion frequently mentioned by the editors is one which, in the eyes of several editors, is the essence of good science. "Theories and rationalizations often don't stand the test of time," one editor remarks, "but the experiment should." Another states that papers based on "shoddy experimental work" are not accepted by his journal under any circumstances. The third criterion concerns the ability of the author to bring together recent developments within a field and to make sense of them. Editors are careful to distinguish this aspect of the research from its originality, or "sheer novelty" value.

Curiously, the criterion mentioned most frequently in the written instructions to reviewers is not one of the three criteria. Ten of the journals call on reviewers to consider elements of style--the organization of ideas, the clarity and conciseness of the prose, and the proper use of English. It could be that editors do not emphasize scope, quality of experimental work, or significance of results because they assume that reviewers will usually take such essentials into account. Or it could be that only criteria that apply to particular journals need to be communicated to reviewers. In the case of the Physical Review Letters, for example, when the editors decided to shift its emphasis from "novelty and timeliness" to "general interest," they published an announcement that explained the change in policy and the reasons behind it (Lazarus 1980). Accordingly, the concerns of editors about style may be an attempt to make reviewers conscious of the specific editorial policies or preferences of chemistry journals.

The Process of Decision-Making

Editors decide to publish a manuscript or not after a judicious weighing of the comments and recommendations of the reviewers. If the opinion of reviewers about a paper is unanimous, editors nearly always act accordingly. More frequently, however, reviewers disagree. In those situations, editors may take a second look at the manuscript to determine whether they can render a deciding opinion. Otherwise, additional opinions may be sought from more reviewers (at the risk of delaying the final decision for months) or from members of an editorial board. In addition, the author(s) may be asked to reply to the criticisms raised by the first set of reviewers. To reach the final decision, editors usually consider the comments in terms of who said what. Thus the opinion of a researcher who is very active and knowledgeable in the field or of a scientist whose past judgments have generally been proved correct may weigh heavily.

The effort expended on papers about which reviewers disagree seems to suggest that the peer system can in fact distinguish differences in scientific quality or merit. This assumption has been examined at least twice in recent years by editors of scientific publications, using data from their files. Cheves Walling (n.d.), the senior editor of the Journal of the American Chemical Society (JACS), looked at pairs of reviewers' recommendations about whether or not to publish, among random samples of 121 "communications" (papers of only a few pages) and 85 longer papers. The distributions of recommendations were compared with distributions that would be expected under various assumptions about the integrity of reviewers and the quality of manuscripts. Walling found that the actual distributions agreed closely with a model in which (1) reviewers are presumed to treat each paper fairly, and (2) papers can be classified on the basis of quality into three types: papers that clearly should be published, those that should not be published, and those that Walling characterized as "marginal," reporting competent but not exceptionally good or apparently significant research. According to the model, approximately one-third of the full-length

papers (31 of 85) would be of the first type, and the remainder would be of the third type; of the 121 "communications," 16 percent (19) would be classified as clearly superior (first type), 7 percent (8) as clearly inferior (second type), and the other 77 percent (94) as marginal in quality (third type).

In a study of Physical Review Letters, Adair and Trigg (1979) examined the recommendations of reviewers for a sample of submitted manuscripts. Approximately one-sixth of the manuscripts received a unanimous recommendation to publish, another sixth received a unanimous recommendation not to publish, and the remaining two-thirds were manuscripts about which the referees (usually two in number) could not agree. The editors concluded that, for the large proportion of papers that do not appear exceptional, the final decision to publish or not depends more on chance, especially in the assigning of reviewers, than on the paper's quality.

Summary and Conclusions

In sum, the replies from our sample of editors indicate that the procedures used to select papers for publication in chemical journals depend heavily on the professional judgments of editors and the reviewing scientists. The performance of journal peer reviews, quantitatively speaking, does not seem to justify the effort: Only one-third of papers that are submitted are in fact excluded, and many of the papers that do appear are not unanimous selections.

Despite the finding, both Walling and the editors of Physics Review Letters offer two strong arguments for maintaining the current procedures or some variation, which can be summarized as follows:

First, it is not clear that any other system would yield a significant improvement in the final result, because the majority of papers cannot be clearly distinguished with respect to quality. The editors of the Physical Review Letters note that "if two-thirds of the papers that we accept were replaced by two-thirds of the papers that we reject, the quality of the journal would not be changed." The editors did consider one proposal that, in the interests of economy, Physical Review Letters not use any system to assess scientific merit and simply open the journal to all papers that meet some

minimal standards or criteria. The proposal was rejected, though, because it would have forced the journal to expand to accomodate the increased number of manuscripts likely to be submitted (at significant cost to the readership), many of which might never have been submitted under a system that employs peer judgment.

Second, and more significantly, peer evaluations apparently permit editors to identify exceptional papers, thus helping to ensure that the best ideas and research are available through the literature to the scientific community and that obviously bad research and incorrect results are not promulgated. Those who benefit most include active researchers faced with the problem of staying abreast of current developments in their field. Whether journals actually fulfill the role as scientific gatekeepers, as Walling has observed, depends on the professional judgment of the editor--and, in particular, on his ability to select responsible and conscientious referees and to determine what weight should be given to the views of each.

The subcommittee wishes to thank the following editors for their cooperation in this survey: Joseph F. Bunnett, Accounts of Chemical Research; Russell F. Christman, and Katherine I. Biggs, Managing Editor, Environmental Science & Technology; Mostafa A. El-Sayed, Journal of Physical Chemistry; Frederick D. Greene, Journal of Organic Chemistry; George H. Morrison, Analytical Chemistry; Hans Neurath, Biochemistry; Robert L. Pigford, I&EC Fundamentals; Philip S. Portoghese, Journal of Medicinal Chemistry; Dietmar Seyferth, Organometallics; J. Willard Stout, Journal of Chemical Physics; Cheves Walling, Journal of the American Chemical Society; and Field H. Winslow, Macromolecules.

References

Adair, R.K., and Trigg, G.L. (1979) Should the Character of Physical Review Letters Be Changed? Physical Review Letters 43 (27): 1969-1974.

Cole, J.R., and Cole. (1973) Social Stratification in Science. Chicago: University of Chicago Press.

Crane, D. (1967) The Gatekeepers of Science: Some Factors Affecting the Selection of Articles for Scientific Journals. American Sociologist 2 (November): 195-201.

Garson, L.R. (1980) Computer-Aided Selection of Reviewers and Manuscript Control. Scholarly Publishing (October): 65-74.

Lazarus, D. (1980) Changes in the Physical Review and Physical Review Letters. Editorial. Physical Review Letters 45 (20): 1605-1606.

Walling, C. (n.d.) The Refereeing of Scientific Manuscripts: Does the Peer System Really Work? Unpublished Paper. Journal of the American Chemical Society.

Ziman, J. (1968) Public Knowledge: The Social Dimensions of Science. Cambridge, England: Cambridge University Press.

Zuckerman, H. and Merton, R.K. (1971) Patterns of Evaluation in Science: Institutionalization, Structure and Functions of the Referee System. Minerva 9 (1): 66-100.

APPENDIX C

STUDIES RELATED TO THE EVALUATION OF SCIENTIFIC RESEARCH

Senta A. Raizen and Lawrence S. Wolfarth

A variety of studies related to the evaluation of basic research have been sponsored or performed by federal agencies. The studies summarized below concentrate largely, but not exclusively, on research supported by the two agencies that make the largest federal investment in basic research--the National Science Foundation (NSF) and the National Institutes of Health (NIH). The studies were selected for their relevance to the subcommittee's exploration of evaluation methods employed by federal agencies; hence, the summaries concentrate on the procedures and criteria used in each study, though results generally are noted as well. The appendix is organized according to the type of investigation represented: Assessments of Agency Evaluation Procedures, Evaluations of Agency Programs, Assessments of Bibliometrics, and Retrospective Studies.

Assessments of Agency Evaluation Procedures

Two recently completed studies examined a number of aspects of the proposal review process at NSF and NIH. Proposal evaluation is relevant to postperformance evaluation of research because previous research performance by a proposer is a key criterion in the review. When past research was supported by earlier grants from the agency, evaluations of a proposal requesting funds for continuation of the work constitutes a form of postperformance evaluations.

Accountability Procedures in NSF and NIH Research Grant Systems (GAO 1981). For this analysis the evaluators

from the General Accounting Office (GAO) selected at random a sample of 25 NIH and 50 NSF grants ending in fiscal year 1978. The sample was drawn from the set of all grants for independent research awarded to faculty members at six major research universities. Data for each grant were collected in the following ways: The contents of the agency's folder--e.g., the original proposal and supplementary materials from the principal investigator(s), comments of reviewers, and the recommendations of program officials--were examined and coded as to what kinds of information had been included. Similar information was gathered for grants to the same investigators immediately preceding the sampled grants and for renewal proposals. Some of the principal investigators, program officials, and reviewers were interviewed about the objectives of the research as they understood it and asked whether the objectives had been accomplished.

In all, 23 of the 25 NIH grants and 27 of the 50 NSF grants had been followed by a request for additional funding. Four of the NIH renewal proposals were rejected outright, and an additional seven were funded at levels less than had been requested. None of the NSF proposals were turned down, though ten investigators did not receive all the funding requested. The GAO concluded that the discrepencies between the agencies in the rate of renewal rejection could be explained by differences in the degree to which agency procedures make investigators accountable for what had been achieved with previous funding: NIH required that all renewal proposals restate the objectives of the initial grant and list the publications that resulted; by contrast, NSF only required evidence that the applicants are competent to carry out the research as proposed. Accordingly, the GAO recommended that NSF change its procedures in order to make renewal applicants more responsible for what had been accomplished with agency funds in the previous grant period.

The GAO invited officials from the six major research universities that administered the grants as well as from NSF and NIH to comment on a draft of its report. The administrators, whose comments have been included in the final report, criticized the design of the study for the following reasons: First, given the small number of grants sampled, differences in the rate of rejection for renewal proposals could be due to an unrepresentative sample rather than to differences in procedures. Second, the GAO failed to consider a number

of other factors that could have contributed to the difference: variation among disciplines in the proportion of proposals that are funded, a potentially significant factor because the NIH grants were drawn from the natural sciences while those in the NSF sample came from both the natural and the behavioral sciences; differences in agency policies and practices for renewal proposals, a factor implicit in the finding that 92 percent of the sampled NIH grants compared with only 54 percent of the NSF grants were followed by renewal requests; and differences in the scientific merit of the renewal proposals submitted to each agency or in the types and amount of funding requested.

Peer Review Procedures in the Selection Among Proposals for Independent Research (Cole et al. 1978). Ten NSF programs with different types of proposal review procedures were selected for assessment. From each program, a random sample of 50 grant applications was chosen, including new research projects that had been funded and some that were not--all submitted during fiscal year 1975. The unit of analysis was the principal investigator; the analysis focused on whether the likelihood of being awarded NSF grants differs for scientists with different professional backgrounds and records.

The following data were collected and then analyzed by regression techniques to establish the effect of the following attributes of scientists on the decision to fund: 1) age of principal investigator, 2) prestige ranking of Ph.D. department (using 1964 ACE rankings), 3) type of present institution (doctorate granting or not), 4) rank of current academic department (using 1969 ACE rankings), 5) academic rank, 6) amount of money applied for in first year of proposed work, 7) number of papers published between 1965 and 1974 (as single author, first author, second author, etc.), 8) citation counts, 9) results of any earlier attempts to gain NSF funding, 10) rating given the current proposal by the NSF program director, 11) the average rating given by external reviewers, 12) type of institution of each reviewer, 13) prestige of each reviewer's department, and 14) geographic location of reviewers and applicant.

A follow-up study (Cole et al. 1981a) looked at the funding decisions for 150 additional proposals--50 proposals submitted to NSF during 1976 to each of three programs, 25 that had been funded and 25 that had been

declined. A new set of reviewers identified by the authors, with help from the Committee on Science and Public Policy of the National Academy of Sciences, was asked to take part in a "blind" review of the proposals (i.e., with clues on the identity of the principal investigator(s) removed) as well as in a standard type of review in which the principal investigator(s) were identified. Data collected were ratings on the proposals: the original NSF ratings, reratings of the same proposals by the second set of experts, and reratings of the same proposals when "blinded." Data on authors and reviewers similar to the data in the initial study also were collected.

The initial study found that in general, the NSF peer review system results in proposals being judged on their own merit and, specifically, that decisions are not biased by the status or position of the applicant(s). The authors found high correlations between funding decisions and the ratings given the proposal by peer reviewers, but low or moderate correlations between funding decisions and the professional status of the applicants or their academic departments. An unexpected result, at least in the context of NSF's requirement that applicants demonstrate their ability to conduct the proposed research, was the low correlation between funding decisions and bibliometrically derived indicators of the impact of previous research.

The follow-up study corroborated the initial analysis in finding no evidence of rating bias in favor of eminent or established scientists among the applicants. Moreover, no evidence of bias was found in the selection of reviewers by NSF program directors. On the average, proposals that received high mean ratings from NSF reviewers received high mean ratings from the second set of reviewers, and no important systematic difference was found for the two sets of reviewers for any of the fields. With respect to blinding, the authors concluded that, since anonymity of established researchers was difficult to accomplish and reviewers found blind proposals more difficult to evaluate, and since there was no clear improvement in the quality of reviews, NSF should not change its procedures in order to conceal the identity of applicants.

On examining the scores given to individual proposals, the authors found that the variation in ratings of the same proposal was greater than the variation in the

average ratings of different proposals. For the proposals ranked in the middle three quintiles on the basis of average scores given by NSF reviewers, evidence suggests that funding decisions for 30 to 44 percent of those proposals would have been reversed had another set of reviewers been used. Reversal rates for the quintiles of highest and lowest scored proposals, however, would have been only 10 and 16 percent, respectively. The Coles (1981b) concluded from the evidence that "the funding of a specific proposal submitted to the NSF is to a significant extent dependent on the applicant's luck in the program director's choice of reviewers . . . ."

This last conclusion has become the subject for much comment in both scientific journals and the general press. In a letter to Science, Singer (1981) argued that attributing reversal rates to the element of chance in reviewer selection would not be reasonable until other possible--and likely--causes of reviewer disagreement have been ruled out, including the possibility that NSF reviewers and the experts selected by the Coles used different criteria or stressed the same criteria to varying degree. In the same letter and in subsequent editorials in Science (Clark 1982) and in The Wall Street Journal (November 23, 1981), it is argued that some reversals are inevitable because the proposals in the middle range received scores that lie very near the cut-off point for funding. As The Wall Street Journal concludes, "(A)ll you can ultimately do about the phenomenon is try to design the panels and application routes to make sure the debates [about which proposals should be funded] occur about reasonable alternatives."

Evaluations of Agency Programs

Most federally sponsored evaluations of research performance have as their aim the assessment of specific agency programs or comparisons of performance resulting from alternative funding patterns for research in a given area. The evaluation method most commonly used is some form of peer review of ongoing work or completed work, often in form of published results. Bibliometric analysis (see below) serves as a supplementary method in some studies and occasionally as the primary evaluation method. The subsection below summarizes major NSF program evaluations; for comparison, some studies assessing extramural

NIH programs are included, as well as procedures for evaluating NIH intramural research. The last two summaries in this section deal with evaluations in two other federal agencies particularly aimed at postperformance assessment of research.

National Science Foundation

Oceanography Program (NSF 1980). The universe sampled for this study was all 95 oceanographic projects completed in 1976 and funded by NSF for at least two years. A random sample of 50 such projects was studied, including some successful and unsuccessful applicants for renewal grants. "Control" projects--to provide a standard of performance--were identified by compiling a list of authors not funded by NSF who had published two or more articles during 1975 to 1978 in journals in which NSF-sponsored oceanographic research had appeared, then selecting 25 of these authors at random.

The papers or abstracts resulting from an NSF grant or published during an analogous time period by one of the 25 non-NSF authors were rated by peer reviewers as to how they compared on a scale ranging from 1 to 100 "to all contemporary basic research projects in oceanography." Each "project" received from one to three reviews. In all, 43 of the 75 project ratings were based on reviews of both abstracts and full papers, 19 on reviews of abstracts only, 13 on reviews of papers only. Publication counts and citation counts were also calculated, and the data were compared with the peer ratings.

Proposal and postgrant peer ratings did not match well except for one of the subdisciplines; nor did proposal ratings match publication and citation measures too closely. NSF program directors appeared to be better predictors than peer reviewers, judging by how closely their ratings of proposals compared to postgrant ratings of papers or abstracts. Renewal proposals that were funded came from projects scoring higher (on the average) in postgrant review than renewal proposals that were denied or projects for which no renewal proposal was submitted. The authors compared the peer ratings of abstracts and publications of the NSF projects with those of the control projects and determined that the quality of research supported by NSF was slightly--though not significantly--

higher. There was little difference between NSF and control projects in the number of citations per publications.

The design of the study has been criticized by the Congressional Reference Service (Knezo 1980) for the following reasons: (1) NSF did not design the control group of 25 non-NSF "projects" to be representative of oceanographic research in general. Presumably the population includes marine scientists whose work appears in nonmarine journals as well as scientists with private firms and universities who do not publish. Yet no effort was made to assess whether the characteristics of the control group were comparable to the population as a whole. (2) Neither did NSF design its control group to be comparable to NSF-funded oceanographic research. On the one hand, the comparison of productivity might have been biased in favor of the control "projects," which had to have published at least two papers in order to be included. On the other hand, the analysis appeared to favor NSF, because control "projects" consisted of the papers of a single author while NSF-sponsored projects sometimes involved multiple researchers publishing separately. (3) Because reviewers were closely connected with NSF's oceanography program and familiar with many of the projects, their evaluations might have been biased. Yet NSF did not attempt to assign reviewers at random or to assess the effects of not doing so. (4) NSF did not assess the validity or reliability of its measures of research quality. In particular, the measure which asked reviewers to compare the quality of the work to "all contemporary projects" was unlikely to yield consistent peer ratings because the standard of reference was so vague. (5) Given the variability in the measures of scientific quality, the samples were too small to provide reliable estimates of differences between NSF-sponsored and other researchers.

Chemistry Program: A Proposed Evaluation Study (NSF 1981). The NSF Office of Audit and Oversight has selected 50 NSF-supported projects and 25 other "projects" in chemistry to be reviewed through peer evaluation of papers, reports, and other publications. The NSF-supported projects are a sample drawn at random from the 214 projects completed in 1976. Each of these projects consists of from 1 to 150 publications, a final report, and optional comments

by principal investigators. The 25 "projects" that are to serve as a control group have been defined by identifying through Chemical Abstracts the keywords that characterize the work in half the NSF projects. Using the same keywords, names of authors not supported by NSF but working in the same areas have been identified, and all publications from the same time period as the NSF projects have been located for 25 of these authors. It is these publications that make up the contents of the control "projects."

NSF states that the sample of control projects "will allow only for a check or calibration of the reviewer's perception of the quality of all contemporary research . . . ." Since some 70 percent of all published research in chemistry does not have NSF support and the overlap between NSF and other federal agency support in the same areas is small, a sample of 25 will not contain enough projects to make comparisons between agencies.

At some time in the future, NSF would like each of the 75 projects to receive three separate reviews rating each on how its publications compare to those of all other current projects in the area. Reviewers will also be asked to classify the work as theoretical, empirical, or facilitative, and as original or derivative. Peer ratings are to be compared with bibliometric data "to determine the extent to which citation analysis can be used as a surrogate for more intensive evaluations."

Materials Research Laboratory Program (Ling et al. 1977, see also Ling and Hand 1980). At the time of the study, there were 20 materials research laboratories (MRL's) that received "core," or institutional, funding from either NSF (which supported 16 MRL's), Department of Energy (2 MRL's), or NASA (2 MRL's). The research performed through institutional funding at these laboratories was compared to project-funded research at 15 universities that did not receive any core funding but obtained the largest amount of NSF project funds of all universities over a six-year period.

The research capability of the institutions and quality of projects was evaluated as follows: The professional status of scientists was assessed qualitatively by a panel of 19 experts and quantitatively on the basis of the number of individual scientific awards received. The degree of overlap in research areas, duration and

turnover of research areas, continuity of funding, and concentration of funding were also determined. Availability of equipment was determined for a subsample of eleven MRL's and nine universities.

Publications were evaluated as the principal product of research. In all, 215 materials science experts reviewed 690 papers, stratified according to type of support and selected characteristics of the author's institution. Characteristics considered in the review included 1) quality of the research work as measured by technical depth and accuracy, 2) degree of innovation, 3) impact on scientific progress, and 4) level of interdisciplinary collaboration--one of the rationales for core funding of MRL's. In addition, citation counts were obtained for the 690 papers and for some 1,609 other papers on materials research published in 1973. Cross-checking of highly cited papers against peer reviews revealed that the number of citations correlated with favorable comments by reviewers.

Productivity was estimated by tabulating achievements or publications and comparing them with an estimate of the administrative costs. Some 403 achievements were assessed by the same 19 experts who had rated the professional status of the researchers. Administrative costs were estimated based on the dollars and the time spent on nonresearch activities.

The evaluators found that eminent researchers were attracted to those laboratories at which they were provided with a great range of sophisticated equipment. The administrative costs of MRL's tended to be lower, primarily because the researchers did not need to invest the time needed by independent researchers to write proposals for project grants. No definitive evidence was found, however, to indicate whether research conducted at MRL's was qualitatively better than that done through project grants. In fact, the only difference in the outcomes of research by the two groups pertained to the type of research conducted: Papers published by MRL's emphasized experimental and engineering-oriented results more than those by independent researchers.

Science Information Activities (Manuel et al. 1977). The study was intended to assess the recent increase in quantity and availability of the information media that serve scientists--books, journals, and data systems. The evaluators used citation counts to estimate the

actual use of NSF-sponsored media by scientists and to assess the effects of publications supported by NSF on specific areas of research.

In all, some 500 papers in the information sciences were selected for analysis from more than 1,000 papers identified by sampling at random 5 papers from each of 15 top journals in the field for each year between 1970 and 1974. Only papers for which the following information was available were included among the 500: title, author(s), name of journal reference, affiliation of first author, funding source, and number of citations to the paper two, five, and ten years after publication. Criteria for assessment included the following: 1) productivity --number of articles published compared to NSF's share of total (federal) funds spent in field, 2) relevance or short-term impact--proportion of articles cited in the second year after publication, 3) significance or long-term impact--proportion of articles cited five or ten years after publication and citation frequency, 4) innovation--earliness (time rank) of publication of articles in field.

The findings with respect to NSF-supported research in the information sciences were: First, NSF-sponsored research is as productive as other research, at least in terms of number of publications per share of federal support. Second, publications from NSF-sponsored research are more relevant--that is, more likely to be cited and to be cited three or more times in the second year after publication. Third, by the fifth and tenth years after publication, papers from NSF-sponsored research are still more likely to be cited, though not to any great extent. Fourth, NSF tends to support established areas of research rather than areas marked by novel developments--a pattern also typical of other federal agencies.

International Decade of Ocean Exploration (Harbridge House et al. 1976). At the time of the Harbridge House study, NSF's contribution to the International Decade of Ocean Exploration (IDOE) consisted of four programs, encompassing 14 large projects and some 200 separate research grants. The reports generated by IDOE at the project level were compared to the reports from projects funded by NSF independent of IDOE. In all, 50 IDOE papers and 50 non-IDOE papers were selected as "most

relevant" to the objectives of the IDOE and non-IDOE projects, and this sample formed the basis for the comparison.

All the papers were reviewed by university teams of faculty members and graduate students. In addition, the principal investigators were interviewed. The objective was to rate the 14 large programs and sampled individual projects as to 1) progress--i.e., achievement of objectives set out by IDOE, 2) the general usefulness of results, 3) uniqueness, 4) scientific quality, and 5) amount of information produced. A separate bibliometric study of IDOE publications and citations also was done.

The authors found that progress toward IDOE objectives by the 14 large projects had been hampered by the fact that funding fell short of initial expectations. Nonetheless, desired research outcomes were accelerated for IDOE research grants intended to synthesize new theories and hypotheses from existing information. IDOE research grants with fairly precise objectives also fulfilled expectations. The quality of IDOE-sponsored publications was generally as good as that of publications from independent grant research, though IDOE results were not being used by the ocean-centered industry. The authors experienced difficulties obtaining current information from some investigators. The study suggests that the informal networks of communication that had developed within the IDOE programs had reduced incentives for investigators to disseminate their findings.

Biome Programs (Battelle 1975). This evaluation covered three of the five programs for large-scale integrated research on biomes funded by NSF as part of the U.S. participation in the International Biological Program. The analysis dealt with 481 reports of research--including 112 oral presentations and 58 unpublished manuscripts--generated by the three biome programs. At the same time, a sample of papers was drawn from two issues of the journal Ecology for each of the years 1967 and 1974 (coinciding with the beginning and the end of the biome research programs); the 63 papers from 1969 and the 47 papers from 1974 provided a standard for assessing the quality of reports from the biome research programs.

Criteria of assessment were: 1) type of paper--public relations (such papers were eliminated in the comparisons with papers from Ecology), methodological, descriptive, analytic, or synthetic, 2) scope, and 3) nature

of research--e.g., degree of coupling of subfields, knowledge transfer, compartmentalization. The papers were categorized by type along the three dimensions, and the frequency of types was tabulated and mapped for the three biome programs and the Ecology papers. The 100 or so papers concerned with developing models were examined to identify progress in weather prediction, ecosystem management, and the testing of theories.

The training and background of scientists participating in the programs were determined from vitae, telephone questionnaires, and ancillary information available to NSF. Institutional affiliation and research contributions were established for individuals, and the number of graduate students and their activities were recorded. Management of the biome research programs was assessed and also compared to that of individually funded grants.

The evaluators found that the publications of the three biome research programs covered a wider range of topics with better balance than the publications from independent research grants. Yet large-scale biome research yielded no major theoretical breakthroughs, nor was it better in quality than that supported by project grants. Attempts to establish comprehensive data banks for individual biomes and to develop complex models of ecosystems--major objectives of the biome research programs--were generally unsuccessful. The failure to achieve program objectives was attributed to the lack of research managers who could integrate the activities and exchange of information among the research units.

The Science Development Program (Drew 1975). The sample covered all 130 doctorate-producing institutions rated by Roose and Andersen in 1970. The institutions included all the recipients of two types of NSF science development (SD) grants (university USD, special SSD) and 65 percent of recipients of the third type (department DSD); institutions that had not received any type of award served as controls. To analyze the effects of SD grants, the institutions were divided into two experimental groups (USD recipients, DSD or SSD recipients) and three control groups (non-recipients that ranked higher than recipients before awards had been made, non-recipients that ranked about the same, and non-recipients that ranked lower). The three science fields that received

the largest share of SD funds (physics, chemistry, and mathematics) were examined in greatest detail; the field of history was also examined as a standard of reference.

One-day site visits were made to 9 USD grant recipients, 1 SSD grant recipient, and 7 DSD grant recipients; 5 institutions not receiving SD funds were also visited. Quantitative analyses included as much longitudinal data as could be obtained for the 15-year period from 1958 through 1972. (The SD grant program was initiated in 1965 and reached its highest level of funding in 1969.) Multiple criteria were used to indicate the level of scientific quality of departments: 1) trends in faculty size, 2) research productivity in terms of publications and citations, 3) characteristics of graduate students, 4) rate of production of new Ph.D.'s, 5) characteristics of the institutions employing the new Ph.D.'s, and 6) amount of research funding attracted from sources outside the university.

The authors determined that SD grants had led to institutional and scientific effects, not all of which were positive. On the plus side, the grants served as catalysts for accelerating the development of the capacity for research in many departments. Science centers were built, larger computers purchased, and libraries expanded. Moreover, SD grants enabled departments to increase the size of faculties; the increase corresponded to an increase in the number of articles published in the most cited journals. SD grants did not affect either the number or quality of graduate students nor did they decrease the probability that first jobs of new doctorates would be with low-ranked departments. On the minus side, the erratic pattern of SD and overall federal support resulted in university funds being shifted in order to complete or maintain improvements in science facilities, often at the direct expense of departments in the humanities and behavioral sciences.

National Institutes of Health: Extramural Research

The Wooldridge Report: Biomedical Science and Its Administration: A Study of the National Insitutes of Health (NIH 1965). In 1962, the year of this study, some 1,100 institutions were receiving NIH support for more than 20,000 separate grants and contracts, with 40 out of the 1,100 institutions sharing three-quarters of NIH extramural funds. Site visits were made to 37 institutions chosen by a dollar-weighted sampling procedure, so

that there was some proportionality between the importance of the institution to the NIH extramural program and the likelihood that it would be chosen. Small institutions were deliberately oversampled, however, so that any size-related difference, if present, could be observed. Regional representation was also built into the sampling design.

At the selected institutions, detailed investigations were made of a total of 240 funded research grants, 125 unsuccessful applications, 105 career development and training grants, and 38 center support and other types of large grants. The procedure for selecting research grants was similar to that for selecting the institutions themselves--that is, based on random selection within intervals, determined by the dollar amounts of support received, so that different-size grants were represented.

A total of 77 scientists and administrators participated in the gathering and evaluation of data, assisted by eight consultants with expert knowledge of specific areas of research. The investigators were grouped into 11 technical panels, nine that examined particular scientific fields and two that dealt with administrative and review procedures. From these panels were drawn teams varying from 4 to 13 people to conduct the institutional site visits. Site visits generally lasted two to three days. Within an institution, members of the teams visited specified individuals and projects and then prepared a report on the institution. More than 600 NIH-funded scientists were visited during the five months allotted to the appraisal, and discussions were held with approximately 150 university administrators who dealt with NIH. After all the site visits were completed, the technical panels reconvened to prepare from the reports on individual institutions a final report on their particular scientific field or special topic.

The reports covered the following subjects: 1) Appropriate level of support for the field, 2) quality and originality of the work supported by NIH, importance of the problems being addressed, gaps between basic research and its application, and proportion of funding going to research as compared to development, 3) questions about proposal review, monitoring, and other management procedures, 4) advantages and disadvantages of traditional research grants, large grants, collaborative research programs, and center funding, and 5) questions related to the training of young scientists.

The scientific panels concluded that the quality of research performed by NIH-supported researchers in universities was generally high, though certain subfields and fields were identified as not receiving sufficient funds to ensure continuing high-quality work. The panel on peer review concluded that current procedures were satisfactory and did not require change. The panel on administrative affairs recommended changes in the organization of NIH in order to improve long-term planning capability. Questions were raised about the need to support intramural research laboratories in light of the evidence that some laboratories were pursuing research already being done at universities.

Unfunded NIH Applications (Carter et al. 1978a). A random sample of 178 investigators was selected from ten representative medical schools. The investigators had applied for but did not receive an NIH research grant in 1970 or 1971. The sample was stratified according to whether the investigator had at least one approved but unfunded application or only disapproved applications, continued to be affiliated with the same medical school, applied for or received an NIH grant subsequently, and maintained an affiliation with a clinical or basic science department. Another criterion was the type of professional degree.

Structured interviews were held with the 126 investigators who could be contacted and agreed to be interviewed. The respondents were queried on such matters as 1) the availability of alternative sources of funds and effort expended in getting funding, 2) effects of negative NIH actions on teaching, equipment, and animal colonies, 3) current research program and support, and 4) current professional status, including present job, patient care responsibilities (for M.D.'s), and administrative duties. Usually two interviewers were present at each interview, which took an average of 40 to 50 minutes.

According to the authors of this report, the primary costs of not funding research can be measured in terms of the "time, efforts, and ideas of productive scientists." The 126 investigators in the final sample had made 156 unfunded applications to NIH for 153 projects: 22 percent of the 153 projects were eventually completed as planned, and significant parts of another 20 percent were completed with modifications, but 43 percent were dropped entirely. One year was the median

period from the NIH decision until other funding was secured, though six investigators reported waiting more than three years for funds. Almost ten percent of the investigators abandoned scientific research completely; another ten percent chose to focus on applied research. Some investigators shifted their research interest to "more popular areas" where funding might be easier to obtain. A few unsuccessful applicants believed that the experience caused them to prepare better proposals or improve the quality of their research, but most did not find any benefits in being denied NIH funding.

Comparison of Large Grants and Research Project Grants (Carter et al. 1978b). The large grants included in the sample represented three different types of funding procedures: program projects, specialized centers, and centers funded through core support. In all, 64 large grants were included, including all such grants for three of the ten NIH institutes. To offer a standard of comparison, a sample of 283 project (R01) grants also was chosen from the same programs as the large grants. Two indicators were used to measure the quality of research of R01 and large grants. The first was an investigator's success in competing for NIH funding. This was established for large grants by checking the rate of approval of R01 applications submitted by investigators who participated in the large grants. The second index was citation counts for journal articles resulting from R01 and large grants, adjusted for field and year of publication. Three other matters related to the rationale for large grants--fields of journal publication (clinical, targeted, or basic science), degree of interdisciplinarity, and involvement of broader segments of the scientific community--also were examined.

The findings indicate that large-grant funding accomplishes many of its goals. Large grants from several institutes successfully promoted interdisciplinary collaboration and collaboration between M.D.'s and Ph.D.'s. Papers resulting from large grants appeared and were cited more frequently in clinical journals. At the same time, the general quality of research produced by large grants may be lower; at least, large-grant investigators were less successful in competing for R01 grants than applicants in general. Moreover, papers from large grants were cited less frequently in journals that report basic research than were papers from R01 grants.

Policy for Biomedical Research (Williams et al. 1978). This report is not an empirical study but an essay on cost-benefit analysis as a method for making decisions about supporting biomedical research. The authors find the method to be of little use because it is difficult to predict the outcomes of research, assign dollar values to the ultimate products, or establish direct cause-and-effect relationships between research dollars and advances in science. Current models of scientific progress and of the processes by which (biomedical) R&D results are transformed into (medical) practice have not proved helpful. They tend to be too simplistic. According to the authors, more research is needed on the reliability of NIH predictions of scientific success, which might include comparing different peer review practices, making predictions about the likely progress of a field and subsequently tracking the accuracy of such predictions, and determining why some projects scored high by reviewers have not produced as expected and why other projects that received relatively low scores proved successful.

An Analysis of Research: Publications Supported by NIH (Narin and Gee 1980). Under contract to NIH, Computer Horizons, Inc., authored a series of reports characterizing the published research supported by all but two of the smaller institutes from 1970 through 1976. Some 600,000 papers published between 1970 and 1977 in 295 major biomedical journals were scanned for acknowledgment of NIH or other support. Papers were also coded as to the type of research a journal generally publishes (clinical or basic biological research) and as to which of 48 subject areas the journal covers. The information was added to a computerized data base containing information on each paper taken from the Science Citation Index (SCI). Approximately 80 percent of NIH-supported publications were covered by SCI and could be included in the analysis.

The subject matter of articles supported by each institute as well as their citations were examined. Citation analysis revealed that seven of the ten most cited papers in cancer research and six of the top ten papers in virology published in 1964 had been supported by the National Cancer Institute. No conclusions were drawn, however, about the overall performance of any institutes.

National Institutes of Health: Intramural Research

In addition to funding research proposals from scientists at universities and independent laboratories, NIH operates several intramural research laboratories and branches. Managers of these units, like corporate executives of industrial laboratories, regard evaluation primarily as a tool for making the best possible use of available resources and talent (Stetten 1981). NIH administrators assess the recent activities and potential contribution of scientists when they are considered for appointment or promotion but pay little attention to day-by-day performance.

Instead, emphasis is placed on assessing research projects. This provides NIH administrators information for improving the intramural research effort in two ways: Problems or limitations of current and proposed research projects are identified, and opportunities for collaboration among researchers are highlighted. The branch and laboratory chiefs--experienced scientists familiar with the projects under their supervision--are expected to recommend improvements or changes in research protocols. Collaboration among employees, which contributes to productivity by reducing unnecessary duplication of research and by getting information to scientists quickly, is promoted at the next level up, by the Scientific Directors of the Institutes.

Evaluations based on externally generated evidence also contribute to the future direction of the intramural programs. Each of the ten institutes (or institute-level divisions) has its own board of scientific counselors, that is, six to eight independent scientists, assisted by expert consultants. Twice each year, each board reviews all research--either in progress or proposed--through formal presentations by individual scientists. The assessment and recommendations of each board are made available to NIH administrators in a formal report.

An assessment of the quality of research produced by the intramural research program of NIH as a whole is derived from the publication record. The large number of papers by NIH intramural scientists published in refereed journals has been taken to demonstrate that the quality of the research carried out by intramural employees is comparable to that of academic scientists.

The results of such checks of publications are seen to obviate the need for a more formal review of the entire intramural program.

Program Evaluations in Other Agencies

A Bibliometric Assessment of Sponsored Research (Riecken et al. 1981). This evaluation of the biomedical research program of the Veterans Administration (VA) started with a review of previous studies of bibliometric indicators, especially evidence about citation rates as an index of the quality of biomedical research. Next, an experi- ment was conducted in which experts ranked journals from 24 different fields according to the impact of the research they report. The rankings of the journals were compared to their scores on a bibliometrically derived index, which was computed by dividing the total number of citations to the journal by the number of articles published in it. In the view of the authors, the agree- ment between the experts and the bibliometric ratings of the journals was sufficently high to warrant using the latter as a surrogate indicator of scientific impact.

There were 2,700 articles in 1971 and 1972 reporting original research supported by the VA. However, the analysis was restricted to the 2,145 articles that appeared in journals covered by the Science Citation Index. Citations were totaled for each article for the two years following publication, which for biomedical articles is usually the period of most frequent citation (see Carter 1974). The average number of citations over the two years was then computed for articles grouped according to the first author's affiliation with a medical school. To establish a standard of comparison, the average number of citations for 1973 and 1974 was calculated for 626 papers published in 1971 and 1972 by NIH intramural scientists and for 698 papers by scientists with NIH extramural research grants. Perhaps not surprisingly, given the high quality of NIH-supported research, the papers by NIH scientists received more citations on the average than those by any subset of VA researchers, though not all the differences were large.

Basic Energy Sciences Program (U.S. Department of Energy 1981). The U.S. Department of Energy (DOE) currently is undertaking an assessment of the Basic Energy Sciences

(BES) program, which supports basic research in the natural sciences and engineering by scientists at universities and independent laboratories. The objectives are to: 1) assess the quality of scientists and of research being supported, 2) evaluate the impact of research on the achievement of DOE's mission 3) examine the balance of support among disciplines and the procedures for selecting, managing, and evaluating projects, and 4) decide whether the projects supported are appropriate for DOE.

Between 11 and 16 projects were randomly selected from each of BES's four smaller research divisions and the laboratory and non-laboratory components of the materials and chemistry divisions. The initial sample was supplemented by including five additional projects that had been classified as "long-term" (more than eight years old) and seven more projects classified as "large dollar" (encompassing at present more than $500,000 or five full-time employees). The total number of projects selected was 125, or about 10 percent of the 1,214 projects supported at the time of the sampling.

A form of peer evaluation is to be the primary method for evaluating individual projects. Three to six scientists are to be assembled at one of several sites for two to three days. Prior to the meetings, the members of the panels are to receive materials from project files and supplementary information from the principal investigators. The schedule calls for a 20-minute presentation by the principal investigators, followed by 30 minutes for questions and answers and a further 60-minute period of review for each project. During the presentation and questioning, the following topics are to be covered: 1) the specific scientific problem or question, 2) the design of the research, including the techniques, equipment, and facilities used, 3) previous, current, and anticipated results, and 4) the impact of the research on the mission of DOE.

In their subsequent review, the evaluating scientists are to assess the quality of the principal investigators, the quality of the research being done, and its impact on the mission of DOE, using several rating scales. In a separate analysis, publication and citation counts for key investigators and citation counts for selected publications are to be compared to a normalized distribution of citations for the field in which the investigators had published so as to establish their

relative performance. The results of the bibliometric analysis then are to be matched against the rankings by the panels in order to identify methodological problems with either type of procedure.

Assessments of Bibliometrics

Several program evaluation studies summarized above use bibliometrics either as a check on peer judgment when that is the primary evaluation method used or, in the case of An Analysis of Research: Publications Supported by NIH 1970-1976 (Narin and Gee 1980) and A Bibliometric Assessment of Sponsored Research (Riecken et al. 1981), as the primary method. The studies summarized below concern the advantages and shortcomings of using bibliometric indicators for evaluation.

A Review of Bibliometric Studies (Narin 1976). Narin reviewed 28 studies in which bibliometric indicators were compared with some other measure of scientific performance. Most of the studies had been sponsored by federal agencies that support basic research in U.S. universities; most were intended to determine whether bibliometrics could supplement or replace peer-based procedures for evaluation. Narin examined 11 studies on individual scientists, in which publication or citation rates were compared with such measures of eminence as awards, academic rank, and institutional and professional affiliation. Ten more of the studies covered indices of departmental quality. In these, departmental publication or citation rates were compared with two rankings of academic departments derived from peer evaluations--the Cartter (1966) and the Roose-Anderson (1970) rankings. Six other studies dealt with indices of publication quality by comparing the citation rates with peer evaluations of the same publications. Narin concluded on the basis of the correlations reported in the studies (Pearson's r) that bibliometric indicators agreed reasonably well with other measures of research quality.

The correlation coefficient r, however, indicates only the relative strength of association between two measures. More informative is an index (R^2) of the actual proportion of the total variation of one measure that corresponds to differences in scores on the other, which can be calculated by squaring the correlation coefficient. Although statisticians differ on the exact

numerical criterion, almost all would agree that a measure with an R^2 less than .6 is not an adequate surrogate. For the studies on individual scientists, R^2s generally fall below the cut-off point, ranging from .25 (r=.5) to .64 (r=.8); for the studies on departmental quality, they range from .49 (r=.7) to .81 (r=.9). In other words, bibliometrics appears more appropriate for evaluating departments, research institutions, or other large aggregates of scientists than for evaluating individual researchers.

The Science Citation Index and Chemistry (DeWitt et al. 1980). Some 2,500 faculty members from the 79 university chemistry departments rated by Roose and Anderson (1970) had published one or more scientific papers at the time of the study; these faculty members provided the sample for the analysis. Citation data were collected in the following way: Publications by academic chemists were identified through a search of the 1966-1970 volumes of the Institute for Scientific Information's Source Index. Citations to those publications were compiled by checking the Science Citation Index for 1968-1972, allowing for the usual two-year lag between the publication of a paper and its formal acknowledgment in the chemistry literature. Some 328,000 citations were found for approximately 33,000 publications. Authors were grouped according to their institution, whether they were NSF grantees or declinees or neither, the program in the NSF Chemistry Division to which their work could be associated, their proclivity for being cited (whether or not their papers were cited more than 500 times in the five-year period), and whether or not they were members of the National Academy of Sciences. Annual rates of publication, of citation, and of citation per publication were compared among the various groups.

The authors found that all three bibliometric indicators--publications, citations, and citations per paper--increased along with increases for other measures of scientific quality, "as long as the data are sufficiently highly aggregated." Small differences between individual scientists or aggregates of a few dozen scientists cannot be regarded as significant, however, because the bibliometric measures tend to be widely distributed. In situations where one wants to evaluate objectively the performance of individual scientists, the authors

conclude, reliance on multiple indicators of scientific achievement, including more than one bibliometric measure, may be most effective.

Peer Review, Citations, and Biomedical Research Policy

Carter (1974). The study deals with two methods of assessing research output--peer review and bibliometric analysis. For peer review, the following questions were examined: Are reviewers influenced by the availability of funding or other institutional factors? What is the relationship between review scores on renewal proposals and the scores that the initial grant proposals received? Have judgments about initial proposals changed over time? Carter found that proposal reviewers and experts requested to review the decisions of the original reviewers tended to agree, indicating the existence of criteria for scientific merit that are shared by reviewers within a field.

With respect to bibliometric analysis, the question examined was whether publication and citation counts could be employed as indicators of research productivity. Citations from the years 1968 to 1972 were compiled for some 5,800 papers appearing between 1966 and 1970. The papers represented the published output of all 51 NIH program grants awarded to medical school faculty in fiscal year 1976 and some 747 NIH project grants. Grants that resulted in one or more of the most cited papers were identified, and the success of grantees in obtaining renewal funding was compared with that of other grantees. Also examined for possible effects on citation rates were the source of NIH funding, the funding procedure, and discrepencies in either years or dollars of support between the proposal request and the actual award. Carter concluded that "the judgments of the peer review process are significantly related to an objective measure of research output derived from citations to articles describing the results of the grant."

Retrospective Studies

These types of studies are generally constructed to trace the relationship of previous research to later intellectual or technological developments in a field. Often the motivation is to demonstrate the impact of research funding by a federal agency; in other studies, the purpose is to demonstrate the role of basic research in fostering industrial or medical applications.

Project HINDSIGHT (Sherwin and Isenson 1969). One of the first major retrospective studies, Project HINDSIGHT was intended by the Department of Defense to assess the contributions of basic research and technological research programs to modern weapons systems. Twenty such weapons systems were selected by panels of technical experts, who also identified a total of 710 unique "R&D events"--that is, key discoveries making the systems possible. The events for a single weapons system were examined by the same panel, who classified them according to whether they resulted from basic scientific research or from technological research. The panels also investigated the origins of the research and the circumstances leading to its subsequent exploitation by weapons technologists.

The authors reported the following major findings: (1) The number of critical events necessary for a new weapons systems is proportional to the increase in sophistication of the new system over its predecessor. (2) More than 85 percent of the developments used in new systems were financed by the DOD. (3) Whether such research is conducted in-house, by industry, at universities, or at other research centers does not affect the likelihood of the research contributing to new weapons. The responsiveness of the research to the interests and needs of weapons technologists is a more important factor. (4) Establishing a straightforward relationship between the costs and benefits of a research program is not feasible because of the nature of scientific and technological research. Critics of Project HINDSIGHT (Comroe 1977) have noted that conclustions (2) and (3) have been interpreted by others to mean that basic research is less important to technological innovation than mission-oriented research. Anticipating such an interpretation, the authors of the HINDSIGHT report were careful to highlight the very limited scope of their research: Only one type of technology had been investigated, and only events contributing to improvements in that technology--not events facilitating its initial development--had been examined.

The Kruytbosch Study. Kruytbosch (1978) used panels of experts to identify significant discoveries in four fields of basic research (astronomy, chemistry, earth sciences, and mathematics) between 1950 and 1976. The scientists associated with the discoveries were then

interviewed and asked whether NSF contributed to the preliminary research, either through project grants or through less direct means such as fellowships or instrumentation grants.

Kruytbosch found that NSF support for the highlighted discoveries increased from 20 percent (13 of 64 discoveries) for the period from 1950 to 1968 to 67 percent (14 of 21) from 1968 to 1976. Both percentages are greater than NSF's share of all federal funds for research during those years. The extent of support for discoveries from both time periods by field ranged from 12 percent (2 of 17) for chemistry to 50 percent (9 of 18) for mathematics. Given the large number of organizations and federal agencies supporting research in chemistry and the relative dependence of theoretical mathematicians on NSF, the discrepency is not surprising. One interesting finding concerns the ability of scientists to predict the consequences of their research. Of the 65 projects for which responses could be analyzed, 43 percent (28) of the investigators had made direct reference in the proposal to anticipated consequences of the research actually realized, another 40 percent (26) had proposed work in the general area but had not specified the innovation, and for 17 percent (11) of the projects, the justification of the work given in the proposal was not related to the innovation. This suggests that a requirement that scientists carry out their research to meet the exact objectives they had originally proposed might have reduced their productivity.

T.R.A.C.E.S. (Illinois Institute of Technology 1968). NSF has also supported retrospective analysis in the hope of producing evidence regarding the effects of science policy changes, for example, whether to emphasize basic or applied research. T.R.A.C.E.S. (for "Technology in Retrospect and Critical Events in Science") was a study that investigated five examples of recently developed technologies and products. The authors concluded that approximately 90 percent of the nonmission research essential to a given innovation had been completed at least ten years prior to its successful development, demonstrating that assessing the contribution of basic research to commercial developments requires looking back at events long past.

Industrial Contributions (NSF 1981). NSF, in an exploratory study, is attempting to trace support by its Chemistry Division for research that has led to outstanding industrial products, i.e., products related to physical chemistry that have received awards between 1953 and 1978 from Industrial Research/Development Magazine. So far, research funded by the chemistry programs has been traced to 62 of the 451 award winners. However, in only seven of those cases did the principal investigator of an NSF Chemistry Division research grant actually win the award; for the remainder, the product receiving the award depended on the intellectual contribution of NSF-supported research carried out by someone other than the award winner and made available either in the form of a publication or patent. The implication is that measuring the results of the research funds invested by NSF or other agencies is meaningful only when such concepts as "contribution" or "impact" are clearly defined.

Retrospective Analysis of NIH Contributions to Medicine and Surgery. Probably the most extensive and well-documented retrospective analysis was performed by Comroe and Dripps (1976). Supported in part by the National Heart and Lung Institute of NIH, the authors attempted to formalize the method through the use of experts. Some 40 physicians and 40 to 50 specialists in each of the two fields were asked to pick the top ten advances in cardiovascular and pulmonary medicine and in surgery. The authors and 140 consultants then identified 137 areas of research essential for the ten advances. About 4,000 published articles related to those areas were examined by consultants and the authors, and the 500 or so articles judged to be the most important received detailed study. The work took several years and in the end substantiated the argument by Comroe and Dripps for the need to support basic research: 62 percent of the key articles screened described basic research; 41 percent were not even clinically oriented. Both this study and Comroe's later (1977) account of critical advances in medicine offer evidence that years, decades, or even centuries may pass from the time a discovery is first made to the time its medical implications are realized.

References

Battelle (1975) Evaluation of Three of the Biome Studies Programs Funded Under the Foundation's International Biological Program (IBP). NSF-C879. Columbus, OH: Battelle Columbus Laboratories.

Carter, G.M. (1974) Peer Review, Citations, and Biomedical Research Policy: NIH Grants to Medical School Faculty. R-1583-HEW. Santa Monica, CA: Rand Corporation.

Carter, G.M., Cooper, W.D., Lai, C.S., and March, D.M.S. (1978a) The Consequences of Unfunded NIH Applications for the Investigator and His Research. R-2229-NIH. Santa Monica, CA: Rand Corporation.

Carter, G.M., Lai, C.S., and Lee, C.L. (1978b) A Comparison of Large Grants and Research Projects Grants Awarded by the National Institutes of Health. Report No. R-2228-1-NIH. Santa Monica, CA: Rand Corporation.

Cartter, A.M. (1966) An Assessment of Quality in Graduate Education. Washington, DC: American Council on Education.

Clark, A.H. (1982) Luck, Merit, and Peer Review. Editorial. Science 215 (January 1): 1.

Cole, J.R., and Cole, S., with the Committee on Science and Public Policy (1981a) Peer Review in the National Science Foundation: Phase II of a Study. Washington, DC: National Academy Press.

Cole, S., Cole, J.R., and Simon, G.A. (1981b) Chance and Consensus in Peer Review. Science 214 (November 20): 881-886.

Cole, S., Rubin, L., and Cole, J.R. (1978) Peer Review in the National Science Foundation: Phase I of a Study. Washington, DC: National Academy of Sciences.

Comroe, J.H., Jr. (1977) Retrospectroscope. Menlo Park, CA: Von Gehr Press.

Comroe, J.H., Jr., and Dripps, R.D. (1977) The Top Ten Clinical Advances In Cardiovascular-Pulmonary Medicine and Surgery, 1945-1975. NIH 78-1522. Washington, DC: Department of Health, Education, and Welfare.

Comroe, J.H., Jr., and Dripps, R.D. (1976) Scientific Basis for the Support of Biomedical Science. Science 192 (April 9): 105-111.

DeWitt, T.W., Nicholson, R.S., and Wilson, M.K. (1979) Science Citation Index and Chemistry. Scientometrics 2 (4): 265-275.

Drew, D.E. (1975) Science Development: An Evaluative Study. Washington, DC: National Academy of Sciences.

General Accounting Office (1981) Better Accountability Procedures Needed in NSF and NIH Research Grant Systems. PAD-81-29. Washington, DC: General Accounting Office.

Harbridge House, Costlow, J.D., Kester, D.R., Manheim, F.T., and Polis, D.S. (1976) Evaluation of the International Decade of Ocean Exploration. Report prepared for the NSF Office of Planning and Resources Management, Washington, DC.

Illinois Institute of Technology (1968) Technology in Retrospect and Critical Events in Science. NSF-C535. Chicago: Illinois Institute of Technology Research Institute.

Knezo, G.J. (1980) Analysis of the NSF Report Entitled Evaluation Studies of the NSF's Oceanography and Chemistry Programs: A Progress Report. Congressional Research Service, Library of Congress, Washington, DC.

Kruytbosch, C. with Papenfuss, S. (1978) Some Social and Organizational Characteristics of Breakthrough Science: An Analysis of Major Innovations in Four Fields of Science, 1950-1976. Paper presented at the IXth World Congress of Sociology, Uppsala, Sweden. Planning and Policy Analysis Office, National Science Foundation.

Ling, J.G., and Hand, M.A. (1980) Federal Funding in Materials Research. Science 209 (September 12): 1203-1207.

Ling, J.G., DeBolt, M.A., Lethi, M.T., Stokes, B.B., and Yerhoeff, J. (1978) Evaluation Study of the Materials Research Laboratory Program. Summary Report. McLean, VA: MITRE-Metrek Division.

Manuel, E.H., Jr., Anderson, R.J., Jr., and Duchin, F. (1977) An Evaluation of the Science Information Activities of the National Science Foundation, 1950-1973. Princeton, NJ: Mathtech, Inc.

Narin, F. (1976) Evaluative Bibliometrics: The Use of Citation Analysis in the Evaluation of Scientific Activity. Cherry Hill, NJ: Computer Horizons, Inc.

Narin, F., and Gee, H.H. (1980) An Analysis of Research Publications Supported by NIH, 1970-1976. Washington, DC: Public Health Service, National Institutes of Health.

National Institutes of Health (1965) Biomedical Science and Its Administration. (Wooldridge Report.) Washington, DC: National Institutes of Health.

National Science Foundation (1981) An Analysis of the NSF Chemistry Program Impact on Industrial Innovations Associated with Chemistry. Draft of a Report.

National Science Foundation (1980) Evaluation Study of NSF's Oceanography Program.

Riecken, H.W., Feldman, J.S., and Zelinger, G. (1981) Citation Analyses for Evaluating A Biomedical Research Program: A Bibliometric Study of VA and NIH Research Output. School of Medicine, University of Pennsylvania.

Roose, K.D., and Anderson, C.J. (1970) A Rating of Graduate Programs. Washington, DC: American Council on Education.

Sherwin, C.W., and Isenson, R.S. (1969) Final Report on Project HINDSIGHT. Washington, DC: Office of the Director of Defense Research and Engineering, Department of Defense.

Singer, I.M. (1981) The Peer Review Question. Letter to the Editor. Science 214 (December 18): 1292-1293.

Stetten, D. (1981) Excerpts from "Intramural NIH--Its Status and Prospects (1976)." In National Research Council. Federal Research on the Biological and Health Effects of Ionizing Radiation. Washington, DC: National Academy Press.

U.S. Department of Energy (1981) Basic Energy Sciences Assessment.

Wall Street Journal (1981) Asides: Roll of the Dice Editorial. (November 23)

Williams, A.P., Carter, G.M., Harman, A.J., Keeler, E.B., Manning, W.G., Jr., Neu, C.R., Pierce, M.L., and Rettig, R.A. (1978) Policy Analysis for Federal Biomedical Research. R-1945-PBRP/RC. Santa Monica, CA: Rand Corporation.

APPENDIX D

VISIT TO THE NATIONAL SCIENCE FOUNDATION

On November 19, 1981, a small group representing the Subcommittee on Postperformance Evaluation of Research spent the day at the National Science Foundation. The purpose of the visit was to inform the subcommittee of current evaluation and management procedures within the Chemistry Division and within NSF as a whole, especially those relating to postperformance evaluation. The subcommittee was represented by Gerald Tape, Associated Universities, and Robert F. Boruch, Northwestern University. In addition, Edel Wasserman of du Pont de Nemours & Company, Inc., attended as the liaison member of the Committee on Chemical Sciences of the National Research Council. Other participants were Senta A. Raizen, the subcommittee's Study Director, and William Spindel, Executive Director of the Committee on Chemical Sciences, National Research Council. Taking part for NSF were Richard S. Nicholson, Director, Division of Chemistry; Arthur F. Findeis, Head, Chemical Synthesis and Analysis Section; program directors and officers representing various NSF chemistry programs; Jerome H. Fregeau, Director, Office of Audit and Oversight; and Harry J. Piccariello, Head, Evaluation Staff.

The following summary records the information conveyed to the subcommittee. It concludes with the subcommittee's findings about the visit.

General Information on the Chemistry Division

The total budget for the eight programs of the Division was $57.7 million in fiscal year 1981. This is

a modest share of all funding for research in chemistry. For instance, in 1979, when the NSF Chemistry Division funded $47.7 million in grants and contracts, total federal funding for basic research in chemistry was $189.2 million. About two-thirds of the Chemistry Division's total money allocation in any one year goes to committed continuation funding--i.e., funding of the second or third year of a three-year grant. (Three-year grants funded on an annual basis are standard throughout the Chemistry Division and general throughout NSF. Even though NSF has the authority to make grant commitments for up to five years, Nicholson, the Division Director, considers that in most cases this is too long a period for a research project to go on without review.)

Each year since 1973, the Division has received between 800 to 850 proposals for basic research in chemistry that require peer review. Of these, over half are "renewal" proposals which request support for research that follows up work done under a preceding grant. On the basis of the submissions and the reviews, about 325 to 350 grants are awarded annually, other than second- or third-year funding. Renewal proposals are more successful than new proposals; for example, in 1979, they represented 57 percent of the proposals submitted but 76 percent of the grants awarded.

The staff of the Chemistry Division consists of the Division Director and Section Heads, and about a dozen professionals. All are chemists with research experience. About half of the chemists are permanent staff; the others rotate, spending one or two years at NSF and then returning to their academic institutions.

Renewal Procedures in the Chemistry Division

Nicholson provided the following information: About 95 percent of grant holders in chemistry who are eligible submit renewal proposals, no matter how many earlier grants an individual may have had. Almost all the renewal requests are for the continuation of work under the preceding grant. Complete departures to a new line of research are rare. The 5 percent of grantees who do not resubmit requests are not followed up, but Nicholson suspects that the majority move on to administrative positions or to other non-research jobs and, therefore, do not apply for renewals. Of the 95 percent

who do, about 85 percent are funded, according to program staff. Of the 4,100 chemists in Ph.D. granting institutions, some 800 are supported by NSF, and about one-half of this group develop long-term stability with respect to research support from NSF. Half of the grants at NSF (and at NIH as well) run five to seven years--i.e., two three-year grant periods.

Until recently, all renewal proposals had been similar in style to new proposals, and the review was handled the same way. When possible, some reviewers (usually two out of five or six) who reviewed the original proposal are also asked to review a renewal proposal. Reviewers are asked to assign an overall rating to the proposal--poor, fair, good, very good, excellent. They also are asked for written evaluations on the quality of the proposed research (including comments about originality and creativity), on the recent research achievement(s) of the principal investigator(s), and on the budget and institutional capability.

For the past two years, the Chemistry Division has used two other renewal procedures as well. "Creativity extensions" are for two additional years of funding, giving a total of five rather than three years of funding without a renewal proposal. Principal investigators are notified during the third year of a grant that they will receive a creativity extension for the next two years. Such extensions are awarded on the basis of judgments by NSF program officials as to the scientists who are most productive, one indicator being the quality and quantity of publications. No more than 10 percent of the grantees that come up for renewal in any single year are given such extensions. The same investigator cannot be awarded two creativity extensions in succession. A renewal proposal and full-scale review are necessary after five years, if the grantee wishes to continue the work.

"Accomplishment-based" renewal proposals were introduced in 1979 as an experiment to determine whether such proposals could reduce the workloads of investigators and reviewers. Investigators seeking grant renewals have an option of submitting either a traditional proposal or an accomplishment-based proposal consisting of no more than four single-spaced pages of text, a list of all publications for the past five years (with reference to NSF funding under the current grant), and as many as six reprints of articles that resulted from the

current NSF grant. Accomplishment-based proposals also are not given twice in succession. Reviewers of such proposals are now being asked to comment on the experimental format separately from their proposal review. According to Division officials, by far the majority of reviewers have been positive in their views, often enthusiastic. The Chemistry Advisory Committee looked into the procedures and comments of reviewers in 1980 and made some minor changes, while endorsing the continuation of the concept. NSF will evaluate the experiment again in 1982 to decide whether to make accomplishment-based renewal proposals a permanent option.

Office of Audit and Oversight

This office, which reports to the NSF Director, carries out evaluations and audits for the agency. Fregeau, who heads the Office of Audit and Oversight (OAO), reported that his staff checks on the work done by program officials during the decision-making stages. OAO selects about a 10 percent sample of grant actions (more recently closer to 5 percent because of staff shortages) to determine whether the actions recommended and taken make procedural sense and whether a reasonable case could be made in supporting the action to a non-specialist. The sampling is not random. Special attention is paid to proposals and grant actions that might have unusual characteristics or involve difficult decisions.

A second oversight function is carried out through the advisory committees to the NSF research divisions. These committees also sample grant actions, reviewing for such elements as the competence and number of reviewers, their possible biases, comments and documentation by program officers, distribution of funds among subspecialties, and recognition of and support to new lines of scientific inquiry. Committees make their reports to the research divisions and to OAO. All grant actions are available for inspection by the committees. Each year, they select about 7.5 percent of reviewed proposals and resulting actions. In 1980, more than 700 grant actions were examined in 36 programs and only four actions were identified in which a good case could be made that there should have been a different decision.

As an example of the evaluation staff's activities, Fregeau summarized the design of a proposed evaluation of chemistry projects. A sample of 50 NSF and 25 non-NSF chemistry projects has been selected, and publications associated with the projects have been assembled (see Appendix C for more detail). The publications are to be rated by peer reviewers on the basis of two criteria: First, given the objective of the work, how good is the product? Second, no matter how technically proficient, was the work worth doing? The NSF projects selected include a number that resulted in renewal proposals, some of which were funded and some of which were declined, so that decisions about renewals could be compared with the peer evaluations of the publications resulting from the previous grants.

Fregeau then discussed NSF's response to the report by the Government Accounting Office (GAO 1981, described further in Appendix C), which criticized NSF's review procedures for renewal proposals on the grounds that not enough attention was paid to accomplishments during the earlier grant. NSF has taken four actions: It will ask principal investigators to distinguish more clearly between long-term objectives of their research and results to be expected under a specific grant. Principal investigators will also be asked to name which parts of their work were or are to be supported by NSF and which parts by other agencies. Third, the investigators are to identify specifically what results were achieved under the preceding grant. (Although review forms explicitly ask reviewers to comment on the principal investigator's previous achievements, many reviewers do not make separate comments in this section when they fill out the form.) Fourth, NSF will send out reviewer comments automatically, with the names of reviewers withheld, as suggested by GAO.

Review of Grant Actions

During their visit, the subcommittee members reviewed proposal and grant folders that had been selected by the staff of the Chemistry Division to represent the following categories:

New Proposals
- Clear-cut grant awards
- Clear-cut declinations

Borderline grant awards
Borderline declinations
Young investigators*
Grant awards
Declinations

Renewal Proposals
Clear-cut grant awards
Clear-cut declinations
Borderline grant awards
Borderline declinations
Young investigators*
Grant awards
Declinations
Creativity extensions
Accomplishment-based renewals
Grant awards
Declinations

Regional Instrumentation Facilities Grants

Closed-out Grants

Based on this inspection and subsequent discussion with the staff of the Chemistry Division, subcommittee members made the following observations:

- Reviewer scores are the single most important factor in making funding decisions. The difficult decisions come at a rating of about 4 ("very good"). There is evidence in the folders that the staff takes much time and care over the decisions at the margin and uses such criteria as: What else is going on in the discipline? What is being funded by other agencies? How narrow is the focus of the proposed research? How much of the same type of work should NSF fund and for how long? For young researchers, a score of 4 is interpreted more favorably than for established researchers, and evidence of some form of recognition such as a Sloan Foundation fellowship or postdoctoral appointment to a high-quality department is considered by program staff as an appropriate analog to track record for established researchers.

*Young principal investigators are defined as those who received the Ph.D. degree seven or fewer years previously.

- Staff have to be knowledgeable about the proposed research and about the reviewers, so that reviewer comments can be interpreted accurately. Knowledge of the principal investigator's work (and progress in case of renewal requests) comes from several sources, such as site visits, discussions during professional meetings, and careful reading of reprints and preprints. Site visits are scheduled to take maximal advantage of scarce travel funds--e.g., in conjunction with professional meetings or other visits to a nearby site. They are therefore somewhat random and not a clear part of the review process. Regarding interpretation of reviewer comments, each NSF program official tries to build up a sort of informal calibration of every reviewer. This is a long-term process, because NSF program officials try to use their reviewers sparingly. For example, in one program not more than four proposals are sent to the same reviewer in the course of a year; another program used 512 different reviewers over a three-year period.

- Documentation of the factors that go into staff recommendations, especially for borderline cases, is not sufficient in all instances to allow an outsider to follow the reasoning adequately.

- Subcommittee members observed several instances where formerly productive principal investigators were turned down when seeking continued support because they did not publish the results of earlier grants. Even some eminent investigators have been denied refunding, though in those cases a longer grace period was sometimes observed. The staff of the Chemistry Division noted that pressure to publish is very great in chemistry; hence, it is rare that a principal investigator with an apparently low rate of publication during an earlier grant gets refunded.

- Subcommittee members found praiseworthy the experiments with renewal procedures but wondered whether the experiments are documented adequately. Though the Division sends newsletters to chairmen of chemistry departments, and an article about its

procedures was published in Chemical and Engineering News (August 18, 1980, pp 19-20), many scientists do not know about the innovations. Plans for assessing the effects of the experiments also seem inadequate.

- It is not clear on what basis decisions are made to change the shares of funding that go to each of the eight programs. What goes into the decision to abandon a particular subfield or to invest in a new area?

Nicholson summarized the ways in which the Division tries to stay current--through rotators, through its Advisory Committee which directly addresses the question of funding distribution among subfields in its three-year oversight reviews and also writes essays every year on future trends, through long-range planning which is done in five-year cycles, and through continuous staff interaction with the field and perusal of the literature. However, according to Nicholson, the system is essentially driven by proposals. If good proposals are not received for work in an emerging area, then NSF cannot move into that area. Shifts between subfields are very much at the margin; budget increments or decrements are generally distributed on a pro rata basis. This is not surprising, given that two-thirds of any one year's funding is previously committed and that there are more good proposals in each subfield than can be funded. The Division Director does retain a small reserve to adjust for an imbalance of good proposals received by the different chemistry programs. Given the modest size of individual chemistry grants, it is probably relatively easy to shift emphasis within a program area when a new line of research develops, but such shifts do not show up in the budget process. In fields that require extensive facilities, like high-energy physics, there is much less internal program flexibility because of the larger investments needed for each research effort.

Conclusions

1. The procedures for reviewing proposals and renewals, making decisions on grant awards, and collecting statistics on operations that were reviewed in the

Chemistry Division appear exemplary, careful, and thorough. The small experiments with alternative procedures are to be commended. They show ingenuity and interest in improving proposal and grant management.

2. More documentation may be warranted for this exemplary process for those who are not intimately familiar with it.

3. An important next step is to establish to what extent the procedures and patterns that characterize the Chemistry Division hold for other divisions in NSF. For example, if the 95 percent rate of renewal requests holds for all of NSF and all such renewal proposals are reviewed, then a postperformance evaluation procedure is already in place for a large majority of individual grants made by NSF.