TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE 45, 189-205 (1994)

Quantitative/Qualitative Federal Research Impact Evaluation Practices

RONALD N. KOSTOFF

ABSTRACT

This paper describes the quantitative and qualitative practice of federal research impact evaluation. Evaluation of research impact is described for three cases: Research selection, where the work has not yet been performed; research review, where work and results are ongoing; and ex-post research assessment, where research has been completed and results can be tracked. Qualitative methods (such as peer review) and quantitative methods (such as cost-benefit analysis and bibliometrics) are described. Although peer review in its broadest sense is the most widely used method in research selection, review, and ex-post assessment, it has its deficiencies, and there is no single method that provides a complete impact evaluation.

Introduction

Research is the pursuit and production of knowledge. To measure the impact of research requires the measurement of knowledge. However, knowledge cannot be measured directly. What can be observed and measured are the *expressions* of knowledge, such as papers, patents, and students educated. Measures of the expressions of knowledge resulting from research must of necessity provide an incomplete picture of the research product. The concluding hypothesis that will permeate the remainder of this paper is that the greater the variety of measures and qualitative processes used to evaluate research impact, the greater is the likelihood of converging to an accurate understanding of the knowledge produced by research [1].

Impact of a research program involves identifying the variety of expressions of knowledge produced, as well as the changes that these expressions effect on a multitude of different potential research targets (other research areas, technology, systems, operations, other organizational missions, education, social structures, etc.). Although some impacts may be tangible (new instruments developed, new research fields stimulated, students trained in new disciplines), many may be intangible (e.g., a designer of equipment may receive new insights from having attended a research seminar), and difficult to identify, much less quantify.

RONALD N. KOSTOFF received a Ph.D. in aerospace and mechanical sciences from Princeton University in 1967. At Bell Labs, he performed technical studies in support of the Office of Manned Space Flight, and economic and financial studies in support of AT&T Headquarters. At the US Department of Energy, he managed the Nuclear Applied Technology Development Division, the Fusion Systems Studies Program, and the Advanced Technology Program. At the Office of Naval Research, he is director of Technical Assessment, and his present interests revolve around improved methods to assess the impact of research.

Address reprint requests to Dr. Ronald N. Kostoff, Office of Naval Research, Arlington, VA 22217-5660.

The views expressed in this paper are solely those of the author and do not represent the views of the Office of Naval Research.

0040-1625/94/$7.00

Evaluation of research impact is further complicated by the different perspectives and motivations of the assessors. The quantitative approaches require interpretation by the assessors, and the qualitative approaches rest on the purely subjective judgments of the assessors. The importance of a research program represents a weighting of its quantitative and qualitative impacts on the different potential targets of research. Yet this weighting is dependent on the multiple perspectives of the assessors, including technical, organizational, and personal perspectives [2], and the interplay among these perspectives is not always obvious. Thus, not only is the impact of the research on each of its potential targets dependent on some unknown function of the multiple perspectives of the assessors, but the value and relative ranking of the targets depends on these multiple perspectives as well. Selection of technical methodologies, measures, and assumptions by the assessors may be driven significantly by organizational and personal motivations.

Understanding and measures of the impact of research are desired by research sponsors at every stage of the research cycle, including research topic identification, research selection, research management and evaluation, and research termination/ transition and ex-post analysis. Research impact evaluations are of potential use to sponsors in: "Deciding whether to continue or end the program or to increase or decrease its budget; changing the program, or its management, to improve the probability of success; altering policies regarding the procurement, conduct, or management of research; and/or, building support with policy makers and other constituencies of the program" [3].

There are many bibliographies containing the large number of methods developed to evaluate research conduct, impact, and benefits [4, 3, 5–14]. A relatively small fraction of the methods are actually used in practice by federal research sponsors and evaluators. Of those used in practice, only a small fraction of the results of impact studies are reported in the published literature, and an even smaller fraction are accepted by the final federal decision makers. Although a number of the methods in practice actually used by federal research sponsors to measure impact will be described in the remainder of this paper, one objective will be to *focus on the strengths and weaknesses of these selected methods, rather than simply provide a "shopping list" of techniques.*

Research Impact Evaluation Techniques

Luukkonen-Grunow [8] and Averch [9] provide summaries of major research evaluation methods used throughout the world. The three main categories, in frequency of usage order, are: *peer review, nonquantitative case study and anecdotal approaches, and quantitative methods.* Specific variants of the qualitative and quantitative methods are described briefly, and selected examples of the more prominent applications in the U.S. are presented.

PEER REVIEW

Peer review of research represents evaluation by experts in the field, and is the method of choice in practice in the U.S. [3, 5]. Its objectives run the gamut from being an efficient resource allocation mechanism to being a credible predictor of the impact of research.

The latter issue of peer review predictability directly affects the credibility of technological forecasting. Although studies have been done relating reviewers' scores on evaluation criteria to *proposal* outcomes [14] (Appendix II), the author is not aware of reported studies that have related peer review scores/rankings of proposals to *downstream impacts* of the research on technology, systems, and operations. This type of study would require an elaborate data tracking system over lengthy time periods, which does not exist today. Thus, the value of peer review as a predictive tool for assessing the impact of research

on an organization's mission (other than research for its own sake) rests on faith more than on hard documented evidence.

Peer review problems include [15]: Partiality of peers to impact the outcome for nontechnical reasons, including the organizational and personal reasons mentioned previously in the discussion on multiple perspectives; an "Old Boy" network to protect established fields; a "Halo" effect for higher likelihood of funding for more visible scientists/departments/institutions; reviewers differ in criteria to assess and interpret; the peer-review process assumes agreement about what good research is, and what are promising opportunities. These potential problems should be considered during the process of selecting research impact assessment approaches.

Another problem with peer review is cost. The true *total* costs of peer review can be considerable but tend to be ignored or understated in most reported cases. Because there are many different types of peer review, it is very difficult to provide a total cost rule-of-thumb for generic peer review. Nevertheless, consider the following illustrative example for an order of magnitude estimate on total peer-review costs.

Assume that an interim peer review is desired of a $1 million/yr program at a laboratory. The review mode of operation will be to bring a panel of experts in the different facets of research to the laboratory site for 2 days, and hear presentations from the principal investigators. Assume that the panel consists of 10 experts in research, technology, mission operations, etc., and that eight principal investigators will present their projects to the panel. Direct expenditures, such as panel per diem and travel costs, would be in the neighborhood of $6000–$8000. Any honoraria would increase this cost. Indirect expenditures, such as total reviewer, presenter, staff, and review audience *time* spent toward the review, would be in the range of $125,000. This would include at least the following: (1) presenter time in preparing background material for reviewers to read before review, preparing the presentation, making dry runs for management, etc. [$40,000 estimate—10 days]; (2) panel member time for reading background material (papers, reports, plans), traveling to review, spending time at meeting, writing report, etc. [$50,000–$60,000 estimate]; (3)agency staff time for identifying and soliciting reviewers, establishing review and coordinating with lab, writing reports, etc. [$10,000 estimate-1 month]; (4) Audience (lab management, other lab personnel, other agency representatives, etc.) time at review [$20,000 estimate].

The main conclusion of this discussion is that for serious panel-type peer reviews, where sufficient expertise is represented on the panels, *total real costs will dominate direct costs.* The major contributor to total costs is the *time* of all the players involved in executing the review. With high quality performers and reviewers, time costs are high, and the total review costs can be a nonnegligible fraction of total program costs, especially for programs that are people intensive rather than hardware intensive.

Many studies related to peer review have been reported in the literature, ranging from the mechanics of conducting a peer review, to examples of peer reviews, to detailed critiques of peer reviews and the process itself. In addition to descriptions of peer reviews and processes contained in the reviews and surveys referenced above, other examples of processes and critiques can be found in Chubin [16], Chubin [17], Cicchetti [18], Cole [19], Cole [20], Cole [21], Cozzens [22], DOD [23], DOE [24], Frazier [25], Hensler [26], and Nicholson [27].

Although these reported studies present the process mechanics, the procedures followed, and the review results, the reader cannot ascertain the *quality* of the review and the results. In practice, procedure and process quality are mildly necessary, but nowhere sufficient, conditions for generating a high-quality peer review. Many useful peer reviews

have been conducted using a broad variety of processes, and although well documented modern processes [24] may contribute to the efficiency of conducting a review, more than process is needed for high quality. There are many intangible factors that enter into a high-quality review, and before examples of reviews are presented, some of the more important factors will be discussed.

High-quality peer reviews require as a minimum the conditions summarized from Ormala [28]: (1) the method, organization, and criteria for an evaluation should be chosen and adjusted to the particular evaluation situation; (2) different levels of evaluation require different evaluation methods; (3) program and project goals are an important consideration when an evaluation study is carried out; (4) the basic motive behind an evaluation and the relationships between an evaluation and decision making should be openly communicated to all the parties involved; (5) the aims of an evaluation should be explicitly formulated; (6) the credibility of an evaluation should always be carefully established; (7) the prerequisites for the effective utilization of evaluation results should be taken into consideration in evaluation design.

Assuming these considerations have been taken into account, *three of the most important intangible factors for a successful peer review are: Motivation, Competence, and Independence.* The review leader's motivation to conduct a technically credible review is the cornerstone of a successful review. The leader selects the reviewers, summarizes their comments, guides the questions and discussions in a panel review, and makes recommendations about whether the proposal should be funded. The quality of a review will never go beyond the competence of the reviewers. Two dimensions of competence that should be considered for a research review are the individual reviewer's technical competence for the subject area, and the competence of the review group as a body to cover the different facets of research issues (other research impacts, technology and mission considerations and impacts, infrastructure, political and social impacts). The quality of a review is limited by the biases and conflicts of the reviewers. The biases and conflicts of the reviewers selected should be known to the leader and to each other.

PEER REVIEW OF PROPOSED PROGRAMS

The two largest federal sponsors of research are the National Institutes of Health (NIH) and the National Science Foundation (NSF) [29]. The NSF peer-review process of research proposals illustrates how potential research impact influences selection of new research areas. In the NSF process, proposals received are assigned to program officers for review. The program officers select external peer reviewers and using mail, panel, or mail and panel approaches, have the proposals assessed and rated. The program officers then perform their own assessment of the proposals and forward their recommendations to higher levels. These recommendations are rarely overturned [25].

From the 1987 version of the NSF Brochure, *Information for Reviewers*, reviewers use four criteria to assess the proposals: research performance competence; intrinsic merit of the research; utility or relevance of the research; and effect of the research on the infrastructure of science and engineering. Research impacts are evaluated through the second, third, and fourth criteria.

The second criterion, intrinsic merit, incorporates impact of the proposed research on other research fields in its definition, and is a measure of the nearer term impact of the proposed research. The third criterion, utility, addresses the extent to which the work could contribute to an extrinsic goal such as a new technology. The fourth criterion, infrastructure, incorporates impact on the nation's research/education/human resource base.

In the NIH process, proposals are sent to an initial peer-review group, composed mainly of active researchers at colleges and universities, where they are reviewed for scientific and technical merit. After receiving a priority rating from the peer reviewers, the proposals are then sent to a statutorily mandated advisory council for a program-relevance review. After the council members recommend action to be taken on the proposals (usually concurrence with the peer group recommendations, but sometimes special action [25]), the institute staff rank the proposals and initiate a funding strategy.

The review criteria established by Public Health Service regulations and provided to the peer reviewers are: significance and originality of the proposal from a scientific and technical point of view; adequacy of the methodology to carry out the research; qualification and experience of the principal investigator and staff; reasonable availability of resources; reasonableness of the proposed budget and duration of the project; and other factors, such as human subjects, animal welfare, and biohazards. It appears that only the first criterion, significance, relates to impact, and can include the relatively near-term impact on allied research fields. Broader impact and relevance issues appear to be the purview of the advisory councils. The council members are asked to assess the quality of the initial scientific review as well as the proposal's relevance to institute research program goals and broader societal health-related matters.

The Office of Naval Research (ONR) does not have formal peer review of individual research grants, but leaves the choice of peer review to its scientific officers. It requires a competitive process among internal Navy organizations (claimants) with external reviewers for those accelerated program proposals (Accelerated Research Initiatives—ARIs) which constitute about 30% of the total ONR program [30, 31, 14, 32]. The claimants that win the competition then go to the technical community (if their charter is extramural) and advertise their areas of interest for proposals, or, if their charter is intramural, perform the work in-house. The following discussion will be restricted to the selection of the ARIs as reported in Kostoff [30].

At the time the Kostoff paper was published, the ARI competition was centralized, and was used to *allocate resources across claimants and science disciplines.* It was felt that allocation of resources among claimants with charters as divergent as basic, applied, university, and Navy laboratory should be driven by policy rather than interclaimant competition. The ARI competition was decentralized to the claimant level, and is now used to *allocate ARI resources across science disciplines within the claimancies.* The same type of competition reported in Kostoff [30] and the same principles of operation are used by the claimants, with some modifications to account for each claimant's unique aspects.

In the process described in Kostoff [30], all the ARIs proposed by the claimants were categorized into areas of similar science. Panels consisting of experts external to ONR were selected to evaluate the proposed ARIs, with each area of similar science being assigned to one panel. To provide some measure of normalization and standardization across panels, the author served as the Chairman of all the panels.

A key component of the process was the use of *mixed levels of reviewers* on the panels to evaluate the different potential impacts of research. The panels included bench-level researchers to address the impact of the proposed research on the field itself; broad research managers to address potential impact on allied research fields; technologists to address potential impact on technology and the potential of the research to transition to higher levels of development; systems specialists to address potential impact on systems and hardware; and operational naval officers to address the potential impact on naval operations. The presence of reviewers with different research target perspectives and levels of understanding on one panel provided a depth and breadth of comprehension of the

different facets of the research impact that could not be achieved by segregating the science and utility components into separate panels and discussions. The interplay among reviewers coming from different perspectives allowed each reviewer to incorporate elements of other perspectives into his decision-making process.

The panels each met for one day to hear presentations by the proposers, provide written evaluations of each proposed ARI, discuss the written results in detail, and provide a prioritized ranking of the proposed programs. The written evaluation consisted of providing numbers for the following scoresheet factors: research merit (RM); research approach (RA); match between resources and objectives (MBRO); balance between experiment and theory (BBET); probability of achieving research objectives (PARO); potential impact on naval needs (PINN); probability of achieving potential impact on naval needs (PAPINN); potential for transition or utility (PTU); phase of research and development (PRD); overall program evaluation (OPE). Also, a short written description of the naval need was requested, as well as comments to support the numerical scores and to discuss any other issues desired by the reviewer. The panel discussion that followed the presentations revolved around the reviewers' comments and scores, and resulted in a panel consensus score for each factor for each proposed ARI.

The factors on the scoresheet relating to potential research impact estimation are RM, PINN, and PTU. The RM criterion incorporates the potential impact of the research, if successful, on allied research areas. Each panel had two bench-level experts in the research area of each proposed ARI, and one or two panel members who had a broad knowledge of all the research areas being presented. In addition, before the panel met, the Chairman would contact four experts in the research area of each proposed ARI who were not on the panel, in order to gain a better understanding of the research being proposed and to bring additional issues to the panel meeting for the panel's consideration. Thus, the research impact issues were well covered by a substantial concentration of expertise, both formal and informal.

The PINN criterion dealt with downstream impact of the proposed research on naval systems and operations. To address this criterion, each panel included at least one active duty Naval officer, and at least one civilian representative of a Navy office responsible for operations, systems development, or evaluation. To cover the breadth of potential naval applications of many of the research proposals, each of the naval needs experts on the panel would have substantial communication with other experts in different aspects of naval needs before the panel convened, and would usually describe the different applications, or lack thereof, to the panel.

The PTU criterion incorporated the potential nearer term impacts of the proposed research. Transition refers to the actual transfer (or conversion, or metamorphosis) of research programs to exploratory development, or perhaps even advanced development, in the Navy. Because successful dissemination of research results or even insertion into the Fleet are not confined to a transition path as described above, allowance must be made for other uses of a successful program's results. Utility refers to other mechanisms by which the results of a successful program would be transmitted to, and used by, the technical community. The research managers, technologists, and systems and operations experts on the panel insured that there was sufficient expertise to address and score this factor credibly.

The reviewers' bottom line score, OPE, represented the one number that the reviewers' felt characterized their view of the overall quality of the program. It was desired to see how the different factors in the reviewers' scoresheets impacted the bottom line score. For four competition years (1987–1990; actual competitions occurred in 1986–1989), the

scoresheet factors remained fairly constant. Over this period of time, 105 new ARIs were proposed. It was felt that sufficient reviewer scores of these proposed ARIs obtained under essentially the same selection criteria and process existed to allow a meaningful analysis of the characteristics and patterns of the proposals. A parametric study was performed, in part to examine the relationship between reviewers' factor scores and the bottom line score [14].

A multiple regression analysis was performed to relate OPE to the other six major factors on the scoresheet (RM, RA, MBRO, BBET, PINN, PTU) and to subsets of these six factors arising from subcategorizations of the database by technical discipline, claimant, etc. Fifteen different parametric variations were made, but only results for the 105 aggregated proposed ARIs are presented here.

PINN did not weigh as heavily in the reviewers' bottom line score as did PTU. The reviewers appeared to weigh nearer-term impact more heavily in their bottom line decisions, as evidenced by the higher correlations of PTU. Because a separate study [14] showed that the bulk of the proposed ARIs was viewed by the reviewers as basic research, and since the (possibly far) downstream naval impact of basic research may not be evident in many cases, it is not surprising that the more identifiable near-term impacts, such as transition to exploratory development or utility of results by other researchers, would weigh more heavily on the reviewers' bottom line decisions than the longer-term impacts.

PEER REVIEW OF EXISTING PROGRAMS

There are many approaches used by research sponsoring organizations to conduct periodic peer reviews to monitor the quality and potential impact of ongoing research [3, 5–8, 11, 22, 24, 28, 30, 33]. This section focuses on selected peer-review approaches that reflect the state of the art in the technical community and pays special emphasis to how research impact is incorporated into the peer-review process. The first case study is the U. S. Department of Energy (DOE) review of its Office of Basic Energy Sciences (BES), and the evolution of that approach into present DOE practice. The second case study focuses on the ONR methods used to review extramural and intramural programs. The third and fourth case studies relate to the annual review of the National Institute of Standards and Technology (NIST) by the National Academy of Sciences (NAS), and the annual review of the DOE national laboratories by the field offices.

In 1981 the DOE performed an assessment of existing projects funded by its office of Basic Energy Sciences [33, 5, 30]. Out of approximately 1200 active projects supported by BES, a randomly selected sample of 129 projects was reviewed by panels of scientific peers. The projects were grouped by areas of similar science, and the reviews were conducted on 40 separate days by 40 separate expert panels, with an average of four members and three projects per panel. The reviewers were, for the most part, bench level scientists independent of the DOE.

The reviewers were asked to rate several factors for each project: team quality (TQ), scientific merit (SM), scientific approach (SA), productivity (P), importance to mission (IM), energy impact (EI), and overall project quality (OPQ). The three evaluation factors on the scoresheet that related to potential research impact were SM, IM, and EI. SM incorporated the potential impact of the research on allied research fields. IM covered the types of ways in which a research project could contribute to the nation's energy needs. EI was the probable impact of the research project on energy development, conservation, or use.

After the scoring by the panels was completed, all possible linear regression models (ranging from six-factors to one-factor) were used to relate the OPQ rating factore (essentially the reviewers' bottom line score on each project) to the other rating factors for the

129 projects. The six-factor model produced a correlation coefficient of 0.89, which meant that the six factors selected constituted the bulk of the considerations which the reviewers used to score the OPQ rating factor. In fact, the best three-factor model derived to predict the OPQ rating factor score, consisting of TQ, SA, and IM, produced correlation coefficients within three percent of the complete six-factor model [33].

An updated version of the BES evaluation approach is used by the DOE Office of Program Analysis to conduct peer review assessments of DOE research and development [24]. Now, after a panel has completed the evaluation of all the projects assigned to it, the members are asked to identify research needs or opportunities available to the DOE research program. With this updated version, DOE initiated in 1992 a detailed review of all projects supported by BES.

Each of ONR's review processes has a major peer evaluation component adapted to meet the particular needs of the organizational unit under review. The two reviews described here are those of ONR's two largest claimants, the Research Programs Department (RPD) and the Naval Research Laboratory (NRL).

The RPD sponsors extramural basic research mainly at universities, and presently consists of 13 divisions organized along science disciplines. Two separate groups contribute to the one day annual review of each division. One group is the Division's Board of Visitors (BOV), which represents academia, industry, and nonONR government. The majority of the BOV are members of the research community, but typically the BOV will include representatives from the technology development community and the operational Navy. The other group contributing to the review is the Research Advisory Board, the senior management of the RPD whose backgrounds span a wide range of scientific disciplines.

For the review, the Division Director overviews the total division, including programs, accomplishments, new opportunities, and management issues. The division's program managers describe their programs in detail, including the impact on science of their accomplishments, potential or ongoing transitions of their programs to development programs, some bibliometric measures such as publications, and potential impacts on the Navy if successful. The reviewers fill out comment sheets, focusing on scientific merit, technical approach, and potential naval impact, and later discuss their findings with the RPD management.

Almost all of the NRL's programs are intramural, and it conducts full spectrum research in 60 task areas. On average, about 20 task areas will be reviewed per year, with four or five of these task areas reviewed using external reviewers, and the remainder reviewed by an internal NRL management group called the Research Advisory Committee (RAC). The external review group represents academia, industry, and nonNRL government. The RAC consists of NRL senior management whose backgrounds span a broad range of science disciplines.

The Coordinator of the task area reviewed by the external panel overviews the task area and investment strategy. Then, the principal investigators of the task area describe their work in detail, including the impact of their science accomplishments on the task area and allied science fields, transitions to more applied categories, bibliometric measures such as publications and presentations, and potential impact of their research on the Navy.

The reviewers fill out comment sheets, focusing on scientific merit, technical approach, and potential naval impact, and afterward visit and review facilities. The reviewers draft a report and meet with ONR management and members of the RAC to present their preliminary findings. The remaining task areas are reviewed in detail by the RAC.

NIST is reviewed annually by two external groups, a general policy and management review, and a detailed technical review. The Visiting Committee on Advanced Technology reviews general policy, organization, budget, and programs of NIST. The Committee submits an annual report [34], which includes reviews of progress in NIST's science, engineering, and technology transfer programs.

The National Academy of Sciences' (NAS) Board on Assessment of NIST Programs performs a detailed technical review [35]. Seventeen panels of reviewers (about 10 people per panel) from industry and academia conduct program reviews based on 2- or 3-day site visits at NIST facilities. The panels address variants of research quality, and because of NIST's unique charter in supporting competitiveness, pay particular attention to technology transfer, industrial coupling, and emerging technologies. Although quantitative indicators of research impact are not addressed in the panels' annual report [35], impacts of the research on technology and competitiveness are addressed extensively. Recommendations for improvement in these impact areas are provided.

The DOE has nine contractor-operated multiprogram laboratories. Each contractor's laboratory management performance is evaluated annually by the DOE Field Office (FO) to which each laboratory is assigned [36]. The FO prepares an appraisal plan for the laboratory, which focuses on laboratory performance in four areas: (1) institutional management performance, which includes different aspects of overall lab management; (2) programmatic performance, which includes R&D achievements; (3) operations support performance, which includes technical functions which support mission objectives; (4) administrative performance, which includes business management functions [36].

In the programmatic performance areas, sources of input include DOE program officials, other agencies having substantial work at the laboratory, and FO program managers. For this annual review, DOE will use information from its own program advisory committees on the adequacy and impact of the laboratory's R&D efforts in relation to the overall DOE program. Furthermore, DOE will use the reports of the scientific peer review committees established by the contractor, which provide an assessment of the quality of the laboratory's R&D programs.

There appears to be no formal requirement for using teams of external reviewers for the technical programs as in the ONR and NIST reviews; rather, most input seems to come from the sponsors. Estimations of research impact appear to derive from the DOE program advisory committees and peer review assessments, which may be reflected in the annual appraisal.

Conclusions on Peer Review

Peer review is the most widely used and generally credible method used to assess the impact of research. Much of the criticism of peer review has arisen from misunderstandings of its accuracy resolution as a measuring instrument. Although a peer review can gain consensus on the projects and proposals that are either outstanding or poor, there will be differences of opinion on the projects and proposals that cover the much wider middle range. For projects or proposals in this middle range, their fate is somewhat more sensitive to the reviewers selected. If a key purpose of a peer review is to ensure that the outstanding projects and proposals are funded or continued, and the poor projects are either terminated or modified strongly, then the capabilities of the peer review instrument are well matched to its requirements.

The methods that were described include criteria that address the impact of research on its own and allied fields, as well as on the mission of the sponsoring organization. The most intensive use of peer review appears to be the NSF/NIH processes for assessing

proposals, and the NAS annual review of NIST. Nearer-term research impacts typically play a more important role in the review outcome than longer-term impacts, but do not have quite the importance of team quality, research approach, or the research merit.

Based on the regression analyses of the ONR reviewers' scores and the BES reviewers' scores, a minimal set of review criteria should include team quality (if the team is known when proposed programs are evaluated), research merit, research approach, and a criterion related to longer-term relevance to the organization's mission. All of the organizations examined for this study made use of these criteria where applicable, and added other criteria that related to how well the program under review impacted unique facets of the organization's mission. More important than the criteria is the dedication of an organization's management to the highest quality objective review, and the associated emplacement of rewards and incentives to encourage quality reviews.

Quantitative Methods

BIBLIOMETRICS

A recent comprehensive review of bibliometrics [37] shows the sparsity of bibliometric studies for research impact evaluation reported by the federal government. The reason for this is due in part to the following problems with publication and citation counts [15, 38, 7]: (1) Publication counts: indicates quantity of output, not quality; nonjournal methods of communication ignored; publication practices vary across fields, journals, employing institutions; choice of a suitable, inclusive database is problematical; undesirable publishing practices (artificially inflated numbers of co-authors, artificially shorter papers) increasing. (2) Citations: intellectual link between citing source and reference article may not always exist; incorrect work may be highly cited; methodological papers among most highly cited; self-citation may artificially inflate citation rates; citations lost in automated searches due to spelling differences and inconsistencies; Science Citation Index (SCI) changes over time; SCI biased in favor of English language journals; same problems as publication counts.

In response to Cawkell's [39] claims that "citation anomalies have little effect—they are like random noise in the presence of strong repetitive signals," MacRoberts [40] stated the federal concerns about bibliometrics eloquently: "When only a fraction of influences are cited, when what is cited is a biased sample of what is used, when influences from the informal level of scientific communication are excluded, when citations are not all the same type, and so on, the 'signal' may be repetitive, but it is also weak, distorted, fragmented, incoherent, filtered, and noisy."

Another reason for limited federal use can be inferred from Narin [41], where studies on the publication and citation distribution functions for individuals are reviewed. The conclusion drawn, from studies such as those of Lotka, Shockley, De Solla Price, and Cole and Cole, is that *very few of the active researchers are producing the heavily cited papers.* How motivated are funding agencies to report these hyperbolic productivity distributions for different programs in the open literature, especially as many questions exist as to the accuracy and completeness of the bibliometric indicators? This conclusion raises the further question of the role actually played by the less productive researchers (as measured by publication and citation counts): is the productivity of the elite somehow dependent on the output of the less influential, or is the role of the less productive members that of maintaining the stability of the research infrastructure and educating future generations of researchers?

Macroscale bibliometric studies characterize science activity at the national, international, and discipline level. The biennial *Science and Engineering Indicators* report [29] tabulates data on characteristics of personnel in science, funds spent, publications and citations by country and field, and many other bibliometric indicators. Another study at the national level was aimed at evaluating the comparative international standing of British science [42]. Using publication counts and citation counts, the authors evaluated scientific output of different countries by technical discipline as a function of time.

There is little evidence that the results from such studies have much influence on policy or decision-making; that is, the allocation of resources. As Martin et al point out in their conclusions, there is potential benefit for a country to understand its position vis-à-vis that of its competitors in different science areas, in order to be able to exploit opportunities that may arise in those areas. However, which indicators are appropriate and how they should impact allocation decisions are open questions.

With the notable exception of the NIH [7], few federal agencies report use of microscale bibliometric studies to evaluate programs and influence research planning in the published literature. The NIH bibliometric-based evaluations included the effectiveness of various research support mechanisms and training programs, the publication performance of the different institutes, the responsiveness of the research programs to their congressional mandate, and the comparative productivity of NIH-sponsored research and similar international programs.

Two recent papers [43, 44] described determination of whether significant relationships existed among major cancer research events, funding mechanisms, and performer locations; compared the quality of research supported by large grants and small grants from the National Institute of Dental Research; evaluated patterns of publication of the NIH intramural programs as a measure of the research performance of NIH; and evaluated quality of research as a function of size of the extramural funding institution. Most of the NIH studies focused on aggregated comparison studies (large grants versus small, large schools versus small schools, domestic versus foreign, etc).

Patent citation analysis has the potential to provide insight to the conversion of science to technology. Much of the federal government support of the development of patent citation analysis was by the NSF (e.g., 45, 46), although *there is little published evidence now of widespread federal use of this capability.* Some recent studies have focused on utilization of patent citation analysis for corporate intelligence and planning purposes [47]. However, as Pavitt [48] cautions, it is not yet clear to what extent the "other publications," cited in patents, reproduce basic or applied research, from universities or corporate laboratories. In addition, a high proportion [Pavitt's estimation] of technology is not patented, because it is kept secret, because it is tacit and noncodifiable art, or because—as in the case of software technology—it is very difficult to protect through patents.

Despite these limitations, bibliometrics may have utility in providing insight into research product dissemination. For example, in a recent series of presentations to large federally funded laboratories [49], the following suite of bibliometric studies was proposed: (1) examine distribution of disciplines in co-authored papers, to see whether the multidisciplinary strengths of the lab are being utilized fully; (2) examine distribution of organizations in co-authored papers, to determine the extent of lab collaboration with universities/industry/other labs and countries; (3) examine nature (basic/applied) of citing journals and other media (patents), to ascertain whether lab's products are reaching the intended customer(s); (4) determine whether the lab has its share of high impact (heavily cited) papers and patents, viewed by some analysts as a requirement for technical

leadership; (5) determine which countries are citing the lab's papers and patents, to see whether there is foreign exploitation of technology and in which disciplines; and (6) identify papers and patents cited by the lab's papers and patents, to ascertain degree of lab's exploitation of foreign and other domestic technology.

Although it was also recommended that the lab compare its output (papers/citations normalized over disciplines) with that of other similar institutions, this quantitative comparison should be approached with great caution. A recent comparative bibliometric analysis of 53 laboratories [50] clustered the labs into six types (Regulation and Control, Project Management, Science Frontier, Service, Devices, Survey), and stated that "comparisons of scientific impacts should be made only with laboratories that are comparable in their primary task and research outputs." The report concluded further that (1) bibliometric indicators and scientific publications are not the only outputs that should be measured, but the other types of outputs differ for different labs; (2) bibliometric indicators are not equally valid across different types of laboratories; and (3) bibliometric indicators are less useful for the evaluation of research laboratories involved in closed publication markets.

In addition, recent studies were performed [49] to track the dissemination of information from accelerated research programs. Key papers (P1) resulting from these programs were identified, then the citing papers for these key papers (P2) were identified, then the next generation of citing papers (P3), which cited P2 were identified, and so on. The breadth of disciplines impacted by the key papers (P1) can be identified from the succeeding generations of citing papers. The type of analysis done so far provided more of a qualitative than quantitative estimation of breadth of impact.

Preliminary results show that some very fundamental papers impact across a wide spectrum of disciplines, whereas some high quality but more narrowly focused research papers impact one main discipline very strongly through succeeding generations of citations. Because of the large amounts of data required for a complete analysis, especially in which highly cited papers and their descendents are concerned, present efforts are focused on methods to reduce data requirements and still retain a credible analysis.

The author's experience in two funding agencies with bibliometric evaluations of existing and completed research programs, especially citation analysis, is mixed. The positive aspect is that, generally, the relatively productive and high impact researchers working *within* similar areas could be identified objectively. Thus, in evaluations of completed programs, the most heavily cited contributors could be identified for each program, and these "heavy hitters" usually produced what was independently judged as the seminal output for the program. The accuracy of these same-field comparisons was improved somewhat in test cases with the use of laborious normalization processes. Comparing citations from similar topic papers in the same journal issue normalized out publication time, journal, and topical differences.

The negative aspect concerned comparisons *across* programs or technical areas. Some fields, such as the biomedical areas, have a large number of working researchers, and use the open literature as a prime means of communication of research results. The average citation levels of papers in these fields will therefore be high. In other fields, the average citations were low, yet the output was considered high quality by a peer-review approach. Had a straight bibliometric comparison been made among these fundamentally different programs in different fields, erroneous conclusions would have been reached about the relative quality of the research in each of the programs.

Obviously, some normalization across fields is required, and there are commercial organizations with access to multi-field data that do normalize bibliometric results. However, it is not clear to the author at what level the data must be disaggregated in order

for the field differences to be accounted for by normalization. For example, if the lowest level of aggregation for which normalization factors exist is at the level of, say, the field of acoustics, does that mean the factors would apply equally well to the subfields underwater acoustics (UA) and atmospheric acoustics (AA). If not, then a comparison of programs in UA and AA would have an inherent built-in bias.

In the author's view, cross-field comparisons are one of the weak links in practical utilization of bibliometric analysis, and much research needs to be done in this area to improve the utility of the technique. Probably the most useful aspect in the author's experience in using bibliometric analyses to supplement peer review has been identifying discrepancies between the peer review results and the bibliometric results, and trying to ascertain the reasons for these discrepancies. The insights gained have been valuable.

CO-OCCURRENCE PHENOMENA

Modern quantitative techniques utilize computer technology extensively, usually supplemented by network analytic approaches, and attempt to integrate disparate fields of research. One class of techniques that tends to focus more on macroscale impacts of research exploits the use of co-occurrence phenomena. In co-occurrence analysis, *phenomena that occur together frequently in some domain are assumed to be related, and the strength of that relationship is assumed to be related to the co-occurrence frequency.* Networks of these co-occurring phenomena are constructed, and then maps of evolving scientific fields are generated using the link-node values of the networks. Using these maps of science structure and evolution, the research policy analyst can develop a deeper understanding of the interrelationships among the different research fields and the impacts of external intervention, and can recommend new directions for more desirable research portfolios.

Little evidence of federal use of these techniques (co-citation, co-word, co-nomination, and co-classification analysis) has been reported in the open literature. However, as computerized databases get larger, and more powerful computer software and hardware become readily available, their utilization in assessing research impact should increase substantially. These techniques, especially co-word, are discussed in more detail in Appendix III of Kostoff [14].

COST-BENEFIT/ECONOMIC ANALYSES

A comprehensive survey examined the application of economic measures to the return on research and development as an investment in individual industries and at the national level [7]. This document concluded that although econometric methods have been useful for tracking private R&D investment within industries, the methods failed to produce consistent and useful results when applied to federal R&D support.

Cost-benefit analysis has limited accuracy when applied to basic research because *of the quality of both the cost and benefit data due to the large uncertainties characteristic of the research process, as well as selection of a credible origin of time for the computations.* As an illustrative example, an incremental cost-benefit analysis was performed on the fusion-fission hybrid [51]. This study ignored fusion hybrid research expenditures before 1980 (sunk costs). For the parameter ranges chosen, it was shown that there was a broad region over which hybrid development could prove cost-effective. However, *had this same analysis been done in 1934 (around the beginning of identifiable basic research for fusion), using the same cost and benefit streams as in the 1983 study plus adding costs incurred between 1934 and 1980 and discounting back to 1934, then the result would have been much different from the 1983 study.*

In the 1983 study, the problem was treated deterministically; uncertainties or probabilities of success of the different parameter values being achieved were not taken into account. *The real problem, which pervades and limits any attempt to perform a cost-benefit analysis on a concept in the basic research stage, was the inherent uncertainty of controlling the fusion process. This translated to the inability to predict the probabilities of success and time and cost schedules for overcoming fundamental plasma research problems (e.g., plasma stabilities and confinement times); no credible methods were available.* Thus, the main value of the cost-benefit approach was to show that the potential existed for positive payoff from the hybrid reactor development, that there was a credible region in parameter space in which controlled fusion development could prove cost effective; *what was missing was the likelihood of achieving that payoff.*

A more recent study weighed the costs of academic research against the benefits realized from the earlier introduction of innovative products and processes due to the academic research [52]. A survey of corporate R&D executives showed that an average of 7 years elapsed between a research finding and commercialization, and that commercialization would have been delayed an average of 8 years without academic research. A cost-benefit analysis using this survey data showed a very high social rate of return resulting from academic research.

However, the data were not validated independently by a document-based type of analysis (such as TRACES or Hindsight, retrospective studies of innovations [53]) of a sample number of the products and processes. The time between the research findings and commercialization is very short compared to the results of Hindsight or the TRACES studies, and is more in line with the lag time between the *end* of basic research and commercialization shown by Hindsight/TRACES. Use of a shorter lag time in the discounting process increases the benefit/cost ratio and the social rate of return. Although the method is innovative, a more objective data source would provide higher confidence in the computed rates of return.

To summarize the quantitative methods section, few federal agencies report use of bibliometrics to evaluate programs and influence research planning in the published literature. Cost-benefit and other economic approaches have been reported in the published literature over the years. The foundation on which these approaches rest needs to be strengthened to improve their credibility. As Averch [54] states, after describing the huge social rates of return to investments in hybrid corn reported by Griliches [55]: "In general, economists compute high social rates-of-return to most kinds of research. The rates, in fact, are usually much higher than those computed for other kinds of public investment. So there is a puzzle as to why research investments do not increase until their marginal return just equals returns from other public investments."

CONCLUSIONS ON RESEARCH IMPACT EVALUATION TECHNIQUES

Two generic types of research impact assessment approaches used by the federal government were described (peer review and quantitative methods). Peer review is the method used most frequently. All methods examined have their unique shortcomings. A fundamental problem is that many research impact targets exist. These include impact on: research field itself, allied research fields, technology, systems, operations, education, etc. The strength of the specific impact of the research on each of these targets and the weighting assigned to the value of the research impact on each of these targets depends on the technical, organizational, and personal perspectives of the reviewers. For example, although research proposal X may have a very strong potential impact on technology Y and a very weak impact on graduate student education, if the evaluators selected for a

particular review are organizationally and personally inclined to assign high importance to graduate student education, then research proposal X will suffer accordingly. The many available dimensions that derive from these different perspectives serve to complicate the evaluation process.

Much of the research evaluation community has come to believe that simultaneous use of many techniques is the preferred approach. However, there is little evidence of multiple technique use by the federal government in impact assessment, especially bibliometrics to support peer review. This area is ripe for exploitation.

A recent study [9] summarizes quite well the use of research impact assessments by the federal government. "Since 1985, no breakthrough methods of any variety have been invented that more definitively reveal the ex post scientific or social value of past research investments . . . the evidence is sparse that there is much payoff to public or private sector R&D administrators from making greater use of them. . . . R&D administrators do use ex post evaluations for political and organizational purposes, for example, to convince sponsors that they are interested in rational decision processes and that they are funding good work. However, the research evaluation literature between 1985–1990 contains very few demonstrations that evaluation makes any difference at all to the critical decisions about the level and allocation of scarce scientific and technical resources."

Finally, this paper has examined different research impact assessment techniques, and their use by the federal government. The approach has been to describe application of the different techniques, and focus on the strengths and weaknesses inherent in the processes. The paper did not address the predictive reliability of the processes, mainly because there is little literature that provides the basis for predicting which research programs/proposals will have the desired downstream impact. For example, the relationship between a proposal's peer review score or a project's bibliometric rating and the downstream impact on an organization's mission is not addressed in published studies. One could raise the question, as many active researchers have, as to whether there is value to any of these assessment techniques, as their predictive value is unknown. The credibility and predictibility of these assessment techniques are ripe topics for research. A long-term tracking system for research product evolution would be required to gather the necessary data, and the system would require agreement and coordination from a number of the larger federal research sponsoring agencies, and perhaps the larger industrial firms as well. Although such a system would not provide absolute answers, as tracking of the informal modes of knowledge communication would be near impossible, it would provide a much better picture of research impact and its predictability than exists now. Having the long-term data would allow controlled studies to be done comparing the different evaluation techniques among agencies or within agencies. With the present state of information storage and processing capabilities, *research product evolution tracking is an idea whose time has come.*

References

1. Irvine, J., and Martin, B. R., *Foresight in Science: Picking the Winners,* Frances Pinter, London, 1984.
2. Linstone, H. A., Multiple Perspectives: Concept, Applications, and User Guidelines, *Systems Practice,* 2(3), (1989).
3. Salasin, J., et al, *The Evaluation of Federal Research Programs,* MITRE Technical Report MTR-80W123, June 1980.
4. Wirt, J. G., et al, *R&D Management: Methods Used by Federal Agencies,* Rand Report No. R-1156-HEW, January 1974.
5. Logsdon, J. M., and Rubin, C. B., An Overview of Federal Research Evaluation Activities, Program of Plicy Studies in Science and Technology Report, George Washington University, April 1985.

6. Kerpelman, L. C., and Fitzsimmons, S. J., Methods for the Strategic Evaluation of Research Programs: The State-of-the Art, and Annotated Bibliography, NSF Contract No. PRA 8400688, Abt Associates Inc., 1985.
7. OTA, *Research Funding as an Investment: Can We Measure the Returns,* U.S. Congress, Office of Technology Assessment, OTA-TM-SET-36 (Washington, DC: U. S. Government Printing Office, April 1986).
8. Luukkonen-Gronow, T., Scientific Research Evaluation: A Review of Methods and Various Contexts of Their Application, *R&D Management,* 17(3), (1987).
9. Averch, H., Policy Uses of 'Evaluation of Research' Literature, *OTA Contractor Report,* July (1990).
10. Hall, D., and Nauda, A., An Interactive Approach for Selecting IR&D Projects, *IEEE Transactions on Engineering Management,* 37(2), (1990).
11. Johnston, R., Project Selection Mechanisms: International Comparisons, *OTA Contractor Report,* July 1990.
12. OTA, *Federally Funded Research: Decisions for a Decade,* U.S. Congress, Office of Technology Assessment, OTA-SET-490, Washington, DC: U.S. Government Printing Office, May 1991.
13. Kostoff, R. N., *A Quantitative Approach to Determining the Impact of Research,* Presented at Twenty-Second Annual Pittsburgh Conference on Modeling and Simulation, May 2-3, 1991b.
14. Kostoff, R. N., Research Impact Assessment, Presented at Third International Conference on Management of Technology, Miami, FL, February 17-21, 1992a.
15. King, J., A Review of Bibliometric and Other Science Indicators and Their Role in Research Evaluation, *Journal of Information Science,* 13, (1987).
16. Chubin, D. E., and Jasanoff, S., eds., *Special Issue on Peer Review, Science, Technology, and Human Values,* 10(3), 1985.
17. Chubin, D. E., and Hackett, E. J., *Peerless Science: Peer Review and U.S. Science Policy,* State University of New York Press, Albany, 1990.
18. Cicchetti, D. V., The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-Disciplinary Investigation, *Behavioral and Brain Sciences,* 14(1), 1991.
19. Cole, S., Rubin, L., and Cole, J., *Peer Review in the National Science Foundation: Phase one of a study,* National Research Council, 1978, NTIS Acc. No. PB83-192161.
20. Cole, J., and Cole, S., *Peer Review in the National Science Foundation: Phase two of a study,* National Research Council, 1981a, NTIS Acc. No. PB82-182130.
21. Cole, S., Cole, J., and Simon, G., Chance and Consensus in Peer Review, *Science,* 214, November 1981b.
22. Cozzens, S. E., Expert Review in Evaluating Programs, *Science and Public Policy,* 14(2), 1987.
23. DOD, *The Department of Defense Report on the Merit Review Process for Competitive Selection of University Research Projects and an Analysis of the Potential for Expanding the Geographic Distribution of Research,* April 1987, DTIC Acc. No. 88419044.
24. DOE, *Procedures for Peer Review Assessments,* Office of Energy Research, Office of Program Analysis, Report No. DOE/ER-0491P, April 1991.
25. Frazier, S. P., University Funding: Information on the Role of Peer Review at NSF and NIH, U.S. General Accounting Office Report No. GAO/RCED-87-87FS, March 1987.
26. Hensler, D. R., *Perceptions of the National Science Foundation Peer Review Process: A Report on a Survey of NSF Reviewers and Applicants,* NSF Report No. NSF 77-33, 1976, NTIS Acc. No. PB-271775.
27. Nicholson, R. S., *Improving Research Through Peer Review,* National Research Council, 1987, NTIS Acc. No. PB88-163571.
28. Ormala, E., Nordic Experiences of the Evaluation of Technical Research and Development, *Research Policy,* 18, (1989).
29. NSF, *Science and Engineering Indicators—1989,* National Science Board Report NSB 89-1, Washington, DC, U.S. Government Printing Office, 1989.
30. Kostoff, R. N., Evaluation of Proposed and Existing Accelerated Research Programs by the Office of Naval Research, *IEEE Transactions of Engineering Management,* 35(4), (1988).
31. Kostoff, R. N., and Stanford, L. B., Program Funding Profiles under Budgetary Constraints, *Research Evaluation,* 1(1), (1991a).
32. Kostoff, R. N., Evaluating Federal R&D in the U. S., in *Assessing R&D Impacts: Method and Practice,* Bozeman, B., and Melkers, J., eds. (Kluwer Academic Publishers, Norwell, MA) 1992b.
33. DOE, *An Assessment of the Basic Energy Sciences Program,* Office of Energy Research, Office of Program Analysis, Report No. DOE/ER-0123, March 1982.
34. NIST, *Annual Report, 1990,* Visiting Committee on Advanced Technology, January 1991a.
35. NIST, *An Assessment of the National Institute of Standards and Technology Programs: FY 1990,* Board on Assessment of NIST Programs, National Research Council, National Academy Press, 1991b.
36. DOE, Multiprogram Laboratory Appraisals, DOE ORDER 5000.2A, September 1988.

37. White, H. D., and McCain, K. W., Bibliometrics, in Williams, M. E. (ed.), *Annual Review of Information Science and Technology,* 24, 1989.
38. Oberski, K., Some Statistical Aspects of Co-citation Cluster Analysis and a Judgement by Physicists, in Van Raan, A. F. J., (ed.), *Handbook of Quantitative Studies of Science and Technology,* Elsevier, North Holland, 1988.
39. Cawkell, A. E., Understanding Science by Analysing its Literature, in Garfield, E., *Essays of an Information Scientist,* Vol. 2, Phila., PA, ISI Press, 1977.
40. MacRoberts, M. H., and MacRoberts, B. R., Problems of Citation Analysis: A Critical Review, *Journal of the American Society for Information Science,* 40(5), (1989).
41. Narin, F., Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity, NSF Report, NTIS Accession No. PB252339/AS, 1976.
42. Martin, B. R., et al, Recent trends in the Output and Impact of British Science, *Science and Public Policy,* 17(1), (1990).
43. Narin, F., Bibliometric Techniques in the Evaluation of Research Programs, *Science and Public Policy,* 14(2), (1987b).
44. Narin, F., The Impact of Different Modes of Research Funding, in Evered, D., and Harnett, S., eds., *The Evaluation of Scientific Research,* Wiley, Chichester, 1989.
45. Carpenter, M. P., Cooper, M., and Narin, F., Linkage Between Basic Research Literature and Patents, *Research Management,* 13(2), (1980).
46. Narin, F., Noma, E., and Perry, R., Patents as Indicators of Corporate Technological Strength, *Research Policy,* 16, (1987a).
47. Narin, F., Patent Citation Indicators in Strategic Planning, Presented at Symposium on Evaluation of Scientific-Technological Performance, Jahrhunderthalle Hoechst, Frankfurt, Germany, October 5, 1990.
48. Pavitt, K., What Makes Basic Research Economically Useful, *Research Policy,* 20, (1991).
49. Kostoff, R. N., Unpublished data, 1992d.
50. Miller, R., The Influence of Primary Task on R&D Laboratory Evaluation: A Comparative Bibliometric Analysis, *R&D Management,* 22(1), (1992)
51. Kostoff, R. N., A Cost/Benefit Analysis of Commercial Fusion-Fission Hybrid Reactor Development, *Journal of Fusion Energy,* 3(2), (1983).
52. Mansfield, E., Academic Research and Industrial Innovation, *Research Policy,* 20, (1991).
53. Kostoff, R. N., Semi-Quantitative Methods for Research Impact Assessment, *Technological Forecasting and Social Science,* 1993a.
54. Anerch, H., The Practice of Research Evaluation in the United States, *Research Evaluation,* 1:2, 1991.
55. Griliches, Z., Research Costs and Social Returns: Hybrid Corn and Related Innovations, *Journal of Political Economy,* 66, (1958).

Received 18 September 1992; revised 6 February 1993