Pergamon

Evaluation and Program Planning, Vol. 20, No. 1, pp. 77–89, 1997
Published by Elsevier Science Ltd. Printed in Great Britain
0149-7189/97 $17.00+0.00

PII: S0149-7189(96)00038-9

THE KNOWLEDGE POOL: MEASUREMENT CHALLENGES IN EVALUATING FUNDAMENTAL RESEARCH PROGRAMS

SUSAN E. COZZENS

National Science Foundation

ABSTRACT

The Government Performance and Results Act (GPRA) requires all U.S. federal agencies to set measurable goals and report on whether they are meeting them. These requirements force a tradeoff for research agencies. Either they focus on short-term, measurable processes in reporting their performance and neglect the long-term benefits that research produces for economy and society, or they seek relief from the measurement requirements of the law. This article reviews the state of the art in performance measures and assessment processes before GPRA was passed, and discusses the difficulties in fitting these practices into the GPRA framework. It offers a simple logic model for research programs that highlights what is measurable and what is not with regard to activities that build a national science and engineering base. Published by Elsevier Science Ltd

In 1993, the U.S. Congress passed the Government Performance and Results Act (GPRA) with overwhelming bipartisan support. Its purpose is to improve the efficiency and effectiveness of government programs by establishing a system to set goals for program performance and to measure results. The act intends to shift how agencies manage programs from an input focus to an emphasis on performance and results. GPRA grew out of several related government management practices, including the trend in state and local governments and at the national government level in several other countries toward use of program goal-setting and performance measurement.[1] Senator William Roth, who first introduced a similar bill in 1990, calls GPRA "the single most important piece of the puzzle" in improving government performance (Roth, 1995).

The network of U.S. federal agencies that support research at first greeted this act with shock and disbelief. In the fundamental research tradition, planning has been anathema, and it was not immediately obvious what kinds of goals research-sponsoring agencies could adopt without trying to predict discoveries that would be made five years hence. Furthermore, while GPRA calls for annual performance goals expressed in quantified and measurable form, most research assessment has been qualitative. The first reaction to GPRA among many observers in the research community, then, was to hope for an exemption. The Office of Management and Budget quickly made clear that this was not in the cards.

Three years later, research agencies are responding to the law, but they are using all the flexibility that GPRA offers, and inventing some as well. The goal of this article is to show why that special effort is necessary, by identifying the fundamental measurement challenges those agencies face in reporting on their performance. I begin by presenting the GPRA requirements, which closely resemble those of similar legislation in other countries. These requirements add up to a simple pre-

The author is currently Director of the Office of Policy Support at the National Science Foundation, on leave from the Department of Science and Technology Studies at Rensselaer Polytechnic Institute. Any opinions, findings, conclusions, or recommendations in this article are the author's alone and do not represent official views of the National Science Foundation. An earlier version of this article was prepared as part of a project on Metrics of Fundamental Science sponsored by the Office of Science and Technology Policy.

Requests for reprints should be addressed to Susan E. Cozzens, Office of Policy Support, National Science Foundation, 4201 Wilson Blvd, Ste. 1285N, Arlington, VA 22230, U.S.A.

[1]GPRA supplements the call for an emphasis on results in the National Performance Review (Gore, 1993). The Chief Financial Officers Act of 1990 also acknowledged the need for more attention to performance measurement.

scription for agencies: set goals, choose indicators that will tell you whether you are reaching them, and report annually using those indicators. If performance is less than you projected, adjust your actions as necessary to assure that you are both moving and moving in the right direction.

The second section of this paper describes program evaluation practices at U.S. research agencies before GPRA was passed, and reviews the strengths and weaknesses of the performance measures used in those contexts for the kind of aggregate performance monitoring GPRA requires. It illustrates an important point: while objective measures of research performance are available, what defines "high performance" on these measures varies among fields of research. The indicators therefore do not aggregate sensibly to the large reporting units on which GPRA attention focuses.

Furthermore, the available performance indicators refer to short-term outputs, not long-term outcomes of research. The third section of this article presents a generic logic model for a fundamental research program, which shows that the longer-term outcomes of research — which make it worthy of public support — are in principle too difficult to link regularly to research agency activities through performance reporting. The reason for this situation is the knowledge pool (Gibbons & Johnston, 1974) — the confluence of ideas and people that mixes together the results of many research activities and turns them into a potent, public resource for problem-solving and change. What the public wants from research are innovation and discovery, but these are interactive products of the elements of that pool. Because these interactive products appear through unpredictable paths and at uneven intervals, tracking what goes into the pool through to those results is an expensive process, and therefore generally limited to case studies. This fundamental challenge suggests that research outcome measures are not, and in fact will not, become available for GPRA-type performance reporting.

Simple output indicators like those discussed in the second section of the article have been used in the past primarily in the context of expert assessment or detailed program evaluation. In that context, assessors and evaluators are chosen because they have a rich understanding of the complex set of interactions that produce long-term outcomes, and they can use that understanding to interpret short-term output indicators. If used in the absence of the wisdom and judgment of those evaluators, however, such indicators could cause serious distortions by distracting attention away from the creative contributions research is intended to make. The ultimate challenge for research agencies under GPRA, then, is to incorporate that wisdom and judgment into the process of assessing and reporting the performance of research activities.

PERFORMANCE CONCEPTS IN GPRA

The Government Performance and Results Act lists several purposes: improving the confidence of the American people in their government by holding Federal agencies accountable for achieving program results; promoting a new focus on results, service quality, and customer satisfaction; and improving Congressional decision making with better information on the effectiveness and efficiency of programs. To achieve these goals, GPRA calls for a consultative, iterative process of strategic planning and assessment of progress. It requires agencies to develop strategic plans, consulting with Congress in the process; prepare annual plans setting performance goals; and report annually to OMB and Congress on actual performance compared to goals. The law attempts to improve program management directly through the process of producing performance goals and measures, and to improve budget allocation by taking performance information into account.

The act establishes some common vocabulary for discussion of program performance. Implicitly, GPRA treats government activities and spending as *inputs* to a chain of activities that eventually produce benefits for the public. Government inputs are intended to produce both short-term *outputs* and longer-term *outcomes*. The act defines an *output measure* as the tabulation, calculation, or recording of activity or effort. An *outcome measure*, as defined in GPRA, is an assessment of the results of a program activity compared to its intended purpose. To use the example given in one legislative report on the act, eligible clients completing a job training program are outputs; an increase in their rate of long-term employment is an outcome (U.S. Senate, 1993, p. 15).

The guidance accompanying the act also explains that "output measures are often intermediate, in that they assess how well a program or operation is being carried out during a particular time period... Output measures in performance plans should emphasize those used by agency officials in day-to-day operations and program management." (U.S. Senate, 1993, p. 32) The report acknowledges that outcome measurement cannot be performed until a program or project reaches a point of maturity, and that it depends on a clear definition of what results are expected.

Outputs and outcomes, then, are the short- and long-term signs of program performance. A *performance goal*, as defined in the act, is the target level of performance expressed as a tangible, measurable objective, against which actual achievement can be compared. For example, a short-term performance goal for a student reading program is to have 2.3 million students receive an average of three additional hours of reading instruction per week during the 1990 school year (U.S. Senate,

1993, p. 32). A *performance indicator* is a particular value or characteristic used to measure output or outcome. In the previous example, the indicator is hours of reading instruction per week. If a performance goal cannot be expressed in an objective and quantifiable form, an alternative descriptive form may be used. But the indicators must provide a basis for comparing actual program results with the established performance goals (U.S. Senate, 1993, p. 45). OMB has informed agencies that while the goals and indicators should be primarily those used by program managers to determine whether the program is achieving its intended objectives, they should also include measures that will be useful to agency heads and other decision makers in framing an assessment of what the program or activity is accomplishing (Panetta, 1994).

GPRA stresses multiple performance indicators, and emphasizes outcome, rather than output, measures of performance. The report on the bill states

> The Committee believes agencies should develop a range of related performance indicators, such as quantity, quality, timeliness, cost, and outcome... While the Committee believes a range of measures is important for program management and should be included in agency performance plans, it also believes that measures of program outcomes, not outputs, are the key set of measures that should be reported to OMB and Congress. (U.S. Senate, 1993, p. 29)

Indicators are always partial, capturing some aspects and not others of the phenomenon of interest. Even a *set* of performance indicators provides only an approximate representation of a program's actual performance.

Under GPRA, each agency reports as a whole on the performance of a limited set of broad "programs." This provision has important implications for the character of performance indicators. Since most research funding agencies in the United States are parts of larger government departments, most of them will be treated as single programs, or parts of programs, for the purposes of this law. So for example, the National Institutes of Health (which spends over $12 billion annually) is likely to be treated as a single program within the Department of Health and Human Services; and the Office of Naval Research, Army Research Laboratory, and Air Force Office of Scientific Research are all likely to fall into a single reporting category within the Department of Defense. Even the National Science Foundation, which as an independent agency will report directly on its own behalf, has divided its $3 billion dollar budget into just four GPRA reporting areas: research, facilities, education, and administration and management. Thus, performance reporting under GPRA is much more highly aggregated than the research projects that agencies spend the bulk of their time choosing and managing. Yet somehow the results of those projects must be aggregated comprehensively, coherently, and sensibly in that reporting.

GPRA mandates that the General Accounting Office monitor and report on the implementation of GPRA, and assigns responsibility for the implementation itself to the Office of Management and Budget. As required in the act, OMB designated GPRA pilot projects early in 1994. The first set of performance plans submitted under the GPRA pilot projects was an important learning experience for OMB. Twenty percent of the plans were exemplars, demonstrating that measurable, quantitative performance goals could be set in advance. Another 20%, however,

> lacked goals or measures with sufficient substance to be useful in managing a program, measuring performance, or in supporting a budget request. Put another way, if this were... 1997 [when the whole government is required to submit plans], little or nothing worthwhile could be salvaged by OMB from agency plans such as these... A repeat of this experience three years hence would be a major blow to successful implementation of GPRA (Groszyck, 1994).

The conclusion from this exercise was that "...the rest of the government needs to be engaged early-on if useful plans are to be forthcoming in 1997." OMB therefore started in the spring and summer of 1995 to ask agencies to produce parts of what they need to respond to GPRA, and in the summer and fall of 1996 asked for what amounted to a full dress rehearsal for the formal GPRA documents due in September 1997.

PERFORMANCE MONITORING AND PROGRAM EVALUATION

Thus GPRA provides agencies with a standard template for performance planning and reporting: Set goals, choose indicators of progress toward those goals, establish baselines for performance on those indicators, and measure performance annually. Use your measurements to indicate whether you are making progress and are headed in the right direction. GPRA distinguishes between this system of annual performance reporting it mandates and another, related activity, program evaluation. GPRA assumes, rather than mandates, that an agency has an active effort underway in program evaluation. In GPRA, program evaluation plays a different role from performance plans and indicators. Program evaluations are more in-depth studies of program results, and are therefore usually done less frequently and more selectively than performance reporting. Program evaluation often develops output and outcome indicators, but interprets them in a descriptive framework. Agencies could draw GPRA summary indicators from among those developed in detailed program evaluation. But because GPRA performance indicators are

aggregated across programs and need to be gathered and reported annually, as a practical matter they cannot reflect as much depth as the data and information used for a full-blown, detailed program evaluation. For example, full cost–benefit analysis for technological programs is expensive, and is generally done only on a case study basis.

Evaluation of government programs has a history of several decades. (See Shadish et al., 1991; Cook & Shadish, 1986; Rist, 1990; Wye, 1992). The evaluation of government research programs, however, has historically not been closely linked to the larger program evaluation stream, but has instead grown up as an independent craft that shares some values and practices with the other tradition. Research program evaluation, like general program evaluation, is a learning process involving both program participants and stakeholders in an in-depth look at how a program is working. It analyzes the objectives, priorities, and customers for the program; examines the structure of the program's portfolio; and considers the costs of the program in relation to its results. Good research program evaluation is done by independent evaluators, and includes assessors with relevant technical expertise and experiences in the type of research being evaluated as well as assessors from outside the research community. It gathers systematic evidence on program performance and relies on multiple lines of evidence to draw its conclusions, which are reported to program managers and participants, other stakeholders, and the public (Cozzens et al., 1994).

U.S. research agencies have generally followed one of two approaches in their evaluations. One is program review by a panel of external experts, always including researchers and sometimes including users of research results as well (see Table 1). For example, since the 1950s, the intramural programs of the National Institute of Standards and Technology have been evaluated with site visits by expert panels organized by a Board of Assessment which is a branch of the National Research Council. In the same spirit, the Office of Energy Research has a highly structured retrospective process of expert assessment at the project level, with panel scoring on pre-set criteria. The scores are aggregated at program level and reported within the agency.

A second approach to research program evaluation relies more extensively on data gathering by external contractors (see Table 2). Such evaluation studies, which draw more directly on the general program evaluation tradition, often use mail or telephone surveys or publication-based indicators, sometimes in combination with expert judgments of various sorts. An example is the National Science Foundation's 1990 mail survey of participants in its Research Experiences for Undergraduates program. Similarly, to assess prospects for collaboration with industry, the National Institute of Dental Research conducted an extensive study in the area of restorative dental materials research, using publication-based indicators, patent indicators, surveys, and case studies. Evaluation studies have been relatively rare, and are concentrated in the fundamental research agencies, NSF and NIH.

To support technical review and evaluation studies, a set of research performance indicators and techniques has been developed over the years. Wherever accountability legislation like GPRA has been put into effect, research-funding organizations have turned to this conventional set for reporting purposes. The next section considers the theoretical and practical problems of using them in the GPRA context.

SUMMARY PERFORMANCE INDICATORS: PROS AND CONS

All available assessment techniques have both strengths and limitations. Many of the indicators that find useful applications in the context of a research program assessment have more severe limitations for use under GPRA. For example, an assessment panel can take into account descriptive analysis of interview data, complex models of program operation, or sophisticated citation analyses. All of these can provide performance-related information to inform an assessment report, but do not match GPRA's requirements for simple performance indictors. Even simple widely-used output indicators are not designed to be aggregated across fields of research, nor to be used to set performance objectives.

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

Quantity of Knowledge

Publications. Publication counts are by far the most widely used metric of knowledge production in research, finding applications from individual evaluation for promotion and tenure at universities to national science indicators. (On literature-based measures in general, see Cozzens, 1989; Narin et al., 1994; Van Raan, 1993.) European evaluations of university units have routinely included publication counts as one type of productivity index for a decade or so. In the British system, researchers asked for these more objective indicators to be included in the evaluation system to counteract arbitrary judgments by parochial peer reviewers in a first round of university evaluations (Martin & Skea,

TABLE 1
EXAMPLES OF PANEL ASSESSMENT PROCESSES IN U.S. RESEARCH AGENCIES IN 1993

Agency/Office	Cycle	Panelists	Method
DOE/Office of Energy Research	Programs assessed at the request of program or higher level management	Mostly academic scientists, some users	Scoring on pre-set criteria; summary by panel chair
NIST/Board of Assessment (NRC)	Annual process; review is selective	Mix of academic and industrial scientists and engineers	Qualitative assessment
DOE/ Superconductivity Program Review	Annual review of entire program	Mix of academic, government laboratory, and industrial researchers	Scoring on pre-set criteria
DOD/Office of Naval Research	All programs on three-year cycle for in-depth review	Mix of researchers with transition experts and naval users	Qualitative assessment plus scoring on pre-set criteria
DOD/Air Force Office of Scientific Research	Every laboratory task reviewed annually	One panel of scientific advisors; one of transition experts and users	Scoring on pre-set criteria
US DA/ Agricultural Research Service	Areas for assessment chosen strategically	Mix of researchers, large farmers, and industrial users	Qualitative assessment

TABLE 2
EXAMPLES OF EVALUATION STUDIES IN U.S. RESEARCH AGENCIES, 1985–1993

Agency and study	Reason for evaluating	Assessors	Method
NIH: National Cancer Institute, Critical Cancer Research study	Determine effectiveness of various funding mechanisms	NCI P & E staff supervised a contractor	Combinationof expert judgment and publication-based indicators
NIH: National Institute of Dental Research, Restorative Dental Materials Research study	Assess prospects for collaboration with industry; test methods	NIDR P & E staff supervised a contractor	Publication-based indicators; patent indicators; surveys; case studies
NIH: Fogarty International Center, International Research Fellowship Program study	To inform program decisions	FIC Evaluation Staff supervised a contractor	Mail survey
NSF: Research Experiences for Undergraduates	Improve program management	NSF evaluation staff supervised a contractor	Mail survey
NSF: Research Opportunities for Women	Study program effectiveness	NSF evaluation staff conducted the study, with contractor help	Telephone survey
NSF: Small Grants for Exploratory Research	To aid in decision about program expansion	NSF evaluation staff gathered the data; an external panel discussed them and made recommendations	Analysis of management information, plus informed judgment

P & E = Planning and Evaluation

1992; Phillimore, 1989).[2] In the United States, publication counts were among the first evaluative indicators assembled at the National Institutes of Health and National Science Foundation.[3] Technical evaluation panels are often given publication lists for the researchers they are evaluating, for example, at the Office of Naval Research and in the evaluation process for intramural research at NIH (NIH, 1994; Kostoff, 1988). Even programs that carry out no other evaluation activities often include number of publications in their lists of achievements, for example, NASA's microgravity program.

The use of publications as a metric of knowledge output has a long and respected history. It rests on a sociological theory that maintains that the norms of science require researchers to share their results with others in order to get credit for them.[4] However, differences in incentive and reward systems among the sciences and among research settings call for modifications in this theory. For university researchers, the norms of publication are undoubtedly strong. For those in other settings, publication may not be encouraged or may even be actively discouraged.

Some disciplines are also more publication oriented than others. Computer scientists, for example, often claim that programs are their major output, rather than publications. Publication counts are also limited in their applicability to cross-field comparisons because the "least publishable unit" varies among fields of research. Earth scientists publish their work in large chunks, incorporating great swaths of data in models and theoretical arguments. Laboratory scientists, from engineers through molecular biologists, may carve out smaller slices from their flow of work to publish. Social scientists may wait and publish a book (Ziman, 1994, p. 104). Collaboration patterns also affect the number of publications that appear in a field. Finally, the use of publication counts as performance indicators may skew the numbers upward, as researchers respond to this reward system.

As an output measure for research programs, then, publication counts may be a reasonable choice if the publication habits of the scientists supported by the program are fairly similar to each other, and if there is a stable core of researchers who work in settings that encourage publication. Programs that choose to use publication counts as indicators often place boundaries around the set of publications they will choose to count. They may, for example, limit the data to papers that appear in peer-reviewed journals, asking investigators to provide this information. Or they may choose only high-impact journals in the field, or only journals indexed in a prominent indexing service with good coverage.[5] Steps like these assure some homogeneity in the units being included in the metric. Even after these caveats and corrective measures have been taken into account, however, there are inherent limitations in how much publication counts can say about the knowledge outputs of research programs. Fundamentally, publication counts are an output measure, and leave out other important characteristics of the growth of knowledge among researchers supported by a program.

Other Output Measures. Other less widely applicable measures of activity or output are also sometimes used with regard to research programs, when they are deemed appropriate by program managers and participants. These include patents, devices, computer programs, and other signs of invention. For research programs, such data are generally treated as a supplement to knowledge output indicators, but not the major indicator. When patenting activity resulting from basic research programs has been examined, the level of activity is often low (Research Corporation, 1982). Since small numbers are relatively unstable, including such a count in an aggregated set of performance indicators for a research agency is a risky strategy. In the context of detailed program evaluation, however, where a richer set of indicators is examined by a more knowledgeable group of people, even small levels of patent activity may be a relevant sign of certain kinds of important connections between research and the marketplace.

Quality of Knowledge

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

[2]The U.K. funding councils recently decided to limit publication lists submitted by universities in resource allocation processes to the four best papers individuals in departments have published over the last three years (see O'Brien, 1994). It is now not the publication counts that figure in the resource allocation decisions, but rather the quality of the best publications, as judged by peers.

[3]NSF sponsored the first handbook in this area, *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity* (Narin et al., 1976). NIH built an extensive bibliometric data base related to its programs in the 1970s and 1980s, reported in a series of institute-by-institute program evaluation reports.

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

[5]The Engineering Research Council (TFR) in Sweden follows this strategy. One Dutch study also focused on the publications that appeared in top-ranked journals, rather than total publications, as an indicator of quality (Rigter, 1986).

researchers supported by a program, and thus indirectly as an indicator of the quality of the knowledge they produce (see NSF, 1987).

Technical Review. Expert review is the most widely used approach in research evaluation, both in the United States and around the world. "The Nordic Model" of research evaluation, pioneered in Sweden and also used frequently in the other Scandinavian countries, uses small panels of international reviewers, who judge the national effort in a narrow field of research based on a week of site visits to the major laboratories (Ormala, 1989). Research managers have valued these visiting panels more for the place they fill in an overall research management system than for their specific results. For example, in a small country, a peer review panel that judges proposals can become ingrown, or be too soft on researchers who are no longer productive. An external panel that looks at the quality of the projects supported, or even the knowledge that an external panel will at some point be convened, can keep national review panels on their toes and strengthen their resolve with regard to weak research teams (Luukkonen & Stahle, 1993). The Swedish research councils, however, have decided in the last few years that they have learned about as much as they can from that system for the time being, and are experimenting with more comprehensive reviews of larger fields (Wiklund, 1994). For over a decade, the European Community has also relied heavily on technical review to evaluate its programs. But as experience has accumulated, the system has come under criticism for relying too heavily on basic researchers to evaluate applied research programs, and the Community is now considering changes (Skoie, 1993).

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

There are known difficulties in structuring and using technical assessments, even for internal program evaluation purposes. There is no entity "quality" which can be measured objectively, as GPRA requires. Quality is a collective perception, and peer review panels have certain well-known limitations as representations of collective perception. In particular, the results of the evaluation are highly dependent on the choice of reviewers (Cole et al., 1981), and cognitive particularism has been demonstrated — that is, biases of reviewers toward work of the type they are doing (Travis & Collins, 1991). The practice of organizing a good technical review is designed to counteract these problems. Discussions of the state of the art in picking reviewers tend to stress first, getting a breadth of competence that matches the program well, and second, getting well-respected people so that the credibility of the review is established beyond doubt. Independence of reviewers is also considered essential for this purpose. Even after care is taken with these matters, however, technical review remains essentially a process of judgment.

Technical judgment processes would encounter additional difficulties if they were used to produce summary performance indicators to respond to GPRA requirements. One is their cost and intrusiveness. Current best practice for technical review involves face-to-face interaction between researchers and reviewers at a fairly detailed technical level. If this method were applied annually to all, or even a sample of, federal research programs, the price in reviewer time alone would be enormous (see Kostoff, 1994), and would surely violate the principle of keeping GPRA implementation costs to a minimum.

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

Quantifying technical ratings to perform comparisons across fields, however, raises as many methodological problems as the equivalent use of publication counts. Given the sensitivity of technical evaluations to the particular set of individuals involved on the review panel, the reliability of ratings from one year to the next

in an annual process would be open to question under any circumstances. In the context of the budget process, however, where the GPRA performance indicators will be reported, the quantitative ratings provided by technical panels may be even less reliable. When technical reviewers are asked to produce ratings that lead to more or less money for their fields, they tend to skew their ratings upward. NIH has experienced this sort of rating inflation with regard to peer review ratings of proposals, and in fact experimented for a time with normalization systems to correct for such biases. Because of this potential problem, it would be risky at best to set baselines (as GPRA requires) or do comparisons between broad scientific programs (for example, among the three research areas NASA is likely to report on) based on technical review ratings alone.The qualifications about technical review raised in this discussion relate only to constructing quantitative summary performance indicators for GPRA. These caveats need to be carefully distinguished from the pros and cons of using descriptive technical information either as part of the summary performance report for agencies, or in a full program assessment process. In both those other applications, expert judgments are considered essential.

Citation Analysis. It is against the background of limitations to quantitative technical review ratings that counts of citations to publications take on an appeal among some research evaluators. Again, a theoretical framework underpins the use of these counts. The same norms of science that call on researchers to provide public access to their results in the form of publication are thought also to demand that those who receive the results repay the originator with citations. Citations from one paper to another are, in this view, a form of intellectual debt-paying (Merton, 1973; Hagstrom, 1965; Storer, 1966). Whether or not one believes this argument in its entirety, it is clear that the conventions of scientific writing indicate that citations should show some relationship of use or dependence between one article and another.

In this understanding, citations are taken to be in essence an unobtrusive form of wide-scale peer review. Certainly, they add some information to a pure publication count, by indicating whether the work represented in the publications is attracting attention from others in the field. Citation analysts carefully use the word *impact*, rather than *quality*, to refer to what citations count, but they point out that citation counts have been shown in many instances to correlate with peer judgments of quality. One study at the level of individual articles, for example, found that citation counts predicted (in the statistical sense) the quality ratings of each of two technical experts better than the experts' ratings predicted each other (Virgo, 1977). For individual scientists, peer ratings showed correlations with citations in the .6–.9 range in psychology, physics, and chemistry, although the correlation dipped as low as .20 in sociology, in a set of studies reviewed in the classic volume *Evaluative bibliometrics* (Narin et al., 1976, pp. 82–121). At the level of university departments, in biology, physics, chemistry, and mathematics, peer rankings and citations showed .67–.69 correlations (Hagstrom, 1971).

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

- A small share of citations are negative. Studies in the 1970s showed that the share was negligible (Chubin & Moitra, 1975), but high levels of citation to the disputed cold fusion results have raised fears again that locally, the influence of negative citations could be strong.
- Citation numbers are highly dependent on field of research, much more so than publication counts. Biochemists, for example, use an average of about thirty references per article, while mathematicians use only about ten. This effect can be normalized in some kinds of analysis, but doing so takes away the simple, intuitive interpretation of citation statistics.
- In many fields, experimental work tends to be cited more frequently than theoretical, and occasional methods papers achieve extremely high levels of perfunctory citation. Citation counts may thus undervalue growth in understanding and over-value sheer experimental activity — just the opposite of what one would hope for them as a measure of knowledge quality.
- Because the *Science Citation Index* includes references only from journals, in fields where books are a major publication outlet (including the social sciences), citations undercount even impact.

The differences in citation patterns in different fields of research rule out their use as aggregate performance indicators if any comparison across fields is to be done — for example, if NSF were required to report performance for each of its seven research directorates. Within a field, however, since the limitations are likely to apply with equal force over time, citation counts may be useful for setting baselines of visibility for aggregates of publications. Comparison groups can also be constructed for any aggregate of publications based on matched journal sets, to show where that set of publications stand in comparison with others in the same fields. These are types of information that technical ratings cannot provide. NSF discipline-based directorates, for example, could determine their baseline citation rates and as a performance goal try to keep fluctuations from those rates within certain limits, or to stay 25% above the average citations for articles in the

same journals. Since citations peak two to three years after publication, citation information may lag the award of grants by only five to six years, much less than the lag for true outcome indicators. Potentially, then, they could provide useful information for research management purposes, and serve as one among several performance indicators.

Mixed Methods. The strengths and weaknesses of peer review and citation counts appear to be complementary, and evaluators generally advocate using the two together for detailed program evaluation purposes. A technical review panel's judgments, for example, can be challenged by requiring it to study and respond to literature-based data on the program being evaluated (Anderson, 1989). Conversely, professional evaluators can incorporate both citation measures and peer ratings into an overall evaluation report (Mitre Corporation, 1978). These combinations have many advantages for program evaluation purposes, where dialog is possible in the assessment process.

Importance of Research Results

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

User Evaluations. From the standpoint of program evaluation, the key question in judging importance is who does the assessing. Next-stage users are often involved in this judgment. When importance is judged with regard to bodies of scientific knowledge, researchers must judge that quality — but not the researchers supported by the program, nor those who chose the projects it supported. Instead, the next-stage users in this case are researchers outside the program, in the areas where the program's work is claimed to have an impact. Agencies that create generic knowledge resources and human capital can in addition identify stakeholder groups for the resources they produce — that is, groups that use the bodies of knowledge and talent pools that the agencies develop, although not the immediate knowledge outputs of specific projects. Such groups can be involved in program assessment processes.

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

The state of the art in research program evaluation has not developed effective ways to translate the descriptive knowledge that users bring to the program evaluation process into performance indicators. Nor has it needed to, since users could be involved alongside technical reviewers in agency assessment practices. Under the GPRA template of quantitative goals and reports, however, next-stage users would need to be treated as the "customers" for a research program and surveyed for their satisfaction. Appropriate survey instruments and samples could undoubtedly be developed. The Army Research Laboratory, for example, includes customer satisfaction ratings in its summary performance indicators, gathering them on a simple customer feedback form sent out with all final project results.

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

Literature-based Tools. Some sophisticated literature-based techniques have been proposed to give strategic overviews and provide background information for judgments of the strategic contributions of program participants (Callon et al., 1986; Small & Garfield, 1985). Even advocates do not claim, however, that such techniques can be used independently, without interpretation by technical experts; and they are in fact so complex that they have rarely been used in practice (Healey et al., 1986). No simple GPRA performance indicators based on these methods suggest themselves.

Summary

From this discussion, it should be abundantly clear that the methods available for examining the results of research programs may be quite reasonable to use in the context of program assessment, where multiple indicators are the rule and knowledgeable people are available to integrate them wisely into an assessment. Cautions and caveats about such use have been discussed in the preceding subsections, and are already embodied in the practice of research program evalu-

ation, particularly in the use of multiple indicators and their combination with technical review.

A frequently voiced fear about GPRA is that it will encourage agencies to measure what is easy and neglect what is important. One can picture the indicators that would fill this description and satisfy the standard GPRA template:

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

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

The key to responding intelligently to GPRA may therefore lie not in the indicators themselves, but in the larger effort in program assessment in which they are embedded. The indicators can provide a bare-bones description of whether the agency is producing the basic expected outputs. But the more detailed information that is needed for general program planning and resource allocation, including descriptive judgments and analysis, still needs to come from the more intensive and interactive processes of program evaluation and assessment.

THE KNOWLEDGE POOL

The comparative advantage of program assessment over broad, summary performance indicators in the research context stems from the character of research goals. Inputs, activities, and (as discussed in the last section) outputs of research programs are tangible and measurable, but outcomes are much less so. Figure 1 provides a tool for demonstrating this point. It is a highly simplified model of the input–output–outcome relationships in a hypothetical funding program for fundamental research. The inputs shown in the scheme are by and large measurable, and indeed might have served in the pre-GPRA era as performance indicators of sorts.

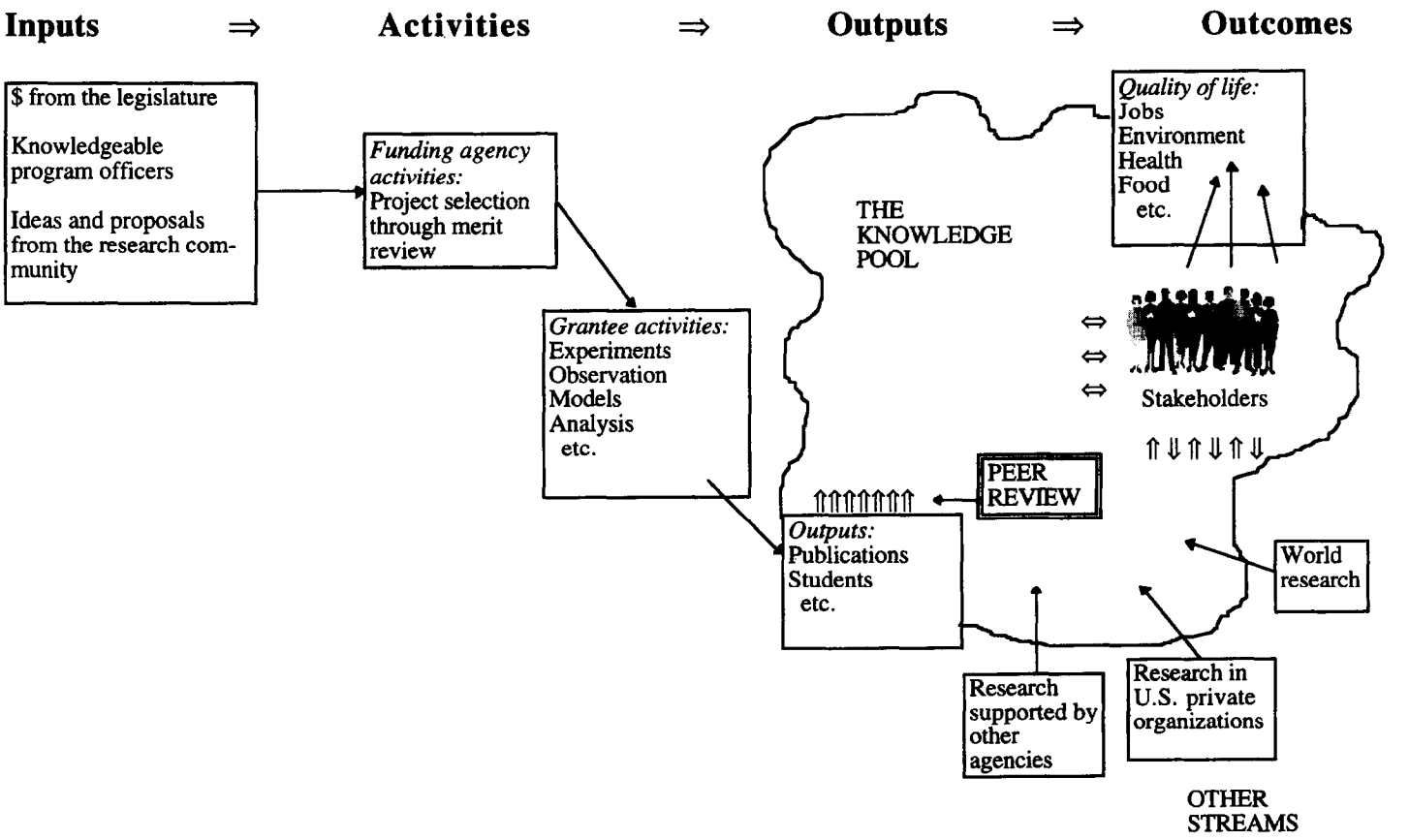

Figure 1. A logic model for fundamental research.

The funding program expends resources ($ from Congress) through administrative processes that require skilled and creative staff, with good access to information. The research community itself provides the raw material of proposals, which put ideas and people at the service of the program. Agencies already have data available in their own records on how many proposals come in and the staff resources devoted to processing them.

When we move into the part of the diagram that represents the selection of projects through the merit review system, we find again that several relevant features can be quantified with agency-held information. For example, agencies can report the number of proposals going through merit review; the number of reviewers consulted and their geographic and demographic distributions; and success rates for the proposals. At NIH, where quantitative rating is the rule in the review system, review scores could be reported. There is general agreement that the available measures of the review process do not reflect its quality or effectiveness very well, but at least activity can be reported. Likewise, estimates of how grantees spend the money are available, for example, how much they requested for salaries, equipment, or student assistance.

But GPRA was passed, not to count agency activities, but rather to define and draw attention to the benefits agencies deliver to the American public. In Figure 1, those benefits lie further down the causal chain. First, the research projects produce outputs. The previous section of this article discussed output measures at length, showing the many methodological pitfalls in aggregating them for agency reporting. But even more fundamental problems in measuring research outcomes appear beyond outputs in the causal chain.

As soon as they are produced, the outputs of research activities join a pool of knowledge and human resources that is fed, not just by one agency's activities, but by the activities of many government agencies, a variety of private organizations such as industrial firms and nonprofit institutions, and the world research community. In the knowledge pool, ideas and people interact and produce innovation and discovery through unpredictable paths and at uneven intervals. The practical value of the knowledge pool is demonstrated concretely only when someone trying to solve a practical problem dips into it for the needed resources — for example, a health professional, a construction engineer, or a government official looking for information. (This group is labeled "stakeholders" in Figure 1, following the terminology introduced in the last section of this article.) The dipping, like the appearance of discoveries, also happens at unpredictable and uneven intervals, and each dip pulls up a mixed product of the many contributing streams. For example, several billion-dollar industries have developed from the interaction of computer science, computer engineering, and commercial development, carried out in universities, government research organizations, and in private industry. Typically, it takes at least fifteen years for commercial products to appear from fundamental advances in the computer field, but the timing and success rate are not predictable (NRC, 1995).

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

How does one measure these long-term, mediated, interactive processes? Research funding organizations can track the outputs of the activities they fund *into* the pool. But if they try to track each drop they have contributed *through* the pool to its outcomes, they will end up spending more money tracking than they spent to support the research. Likewise, when we look back in time from the vantage point of pool-derived innovations or contributions to quality of life, the mixing of streams makes it difficult if not impossible to quantify the contributions of the various sources. Again, unless the goal is specifically to understand the linkages, the expense of doing such an exercise usually outweighs the information gained. For regular performance reporting, such quantified retrospective studies are clearly too expensive and burdensome.

A more reasonable strategy for testing the value of research knowledge to the public relies on human judg-

ment rather than quantitative indicators. Active researchers and stakeholders in a given body of knowledge are aware of the complex processes through which research creates outcomes. They can therefore interpret activity and output indicators in context, and make reasonable judgments of the effectiveness of the funding process in contributing to desirable interactions. As we have seen in previous sections, most research agencies had assessment processes in place even before the passage of GPRA that allowed relevant technical experts and stakeholders to examine the effectiveness of their programs. The challenge for research agencies under GPRA, then, is to develop a reporting system that draws on and encourages the further development of these existing sensible systems.

To accomplish this, many agencies are now turning to the alternative goal-setting and reporting format available under GPRA. This alternative format allows descriptive performance goals, requiring only that agencies paint a word picture of the difference between minimally effective and successful performance. The match between actual performance and the word picture is subject to the same requirements for evidence that apply to quantitative goals. This path thus provides both the flexibility research agencies need to maintain a focus on outcomes rather than outputs of their process, and the rigorous standards GPRA encourages.

CONCLUSIONS

Many of the key issues with regard to implementation of GPRA lie outside the control of agencies, and in the hands of those who receive and use the performance measures. Optimists about GPRA claim that it will revolutionize government management by focusing agency attention diligently on results. Pessimists fear that it will create busywork number-generating, then put a simple-minded tool in the hands of decision makers who already pay too little attention to the programs they expand, cut, and re-arrange. Both optimists and pessimists would probably find some reinforcement for their views in a recent speech by Senator Roth, GPRA's sponsor:

> Imagine what you could do if you combined the kind of program performance information envisioned by GPRA with... program cost-accounting information. We could track the cost-per-unit of activity, and the results of the activity... We could have a sophisticated pay-for-performance system that said, "If you achieve all of your program's managerial goals, and do it under-budget, you will get a significant bonus out of the savings you have created" (Roth, 1995, p.6).

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

REFERENCES

Anderson, J. (1989). *New approaches to evaluation in U.K. funding agencies*, SPSG Concept Paper No. 9. Science Policy Support Group, London.

Callon, M., Callon, Law, J., & Rip, A. (Eds.). (1986). *Mapping the dynamics of science and technology*. London: Macmillan.

Chubin, D. E., & Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science, 5*, 423–441.

Cole, S., Cole, J. R., & Simon, G.A. (1981). Chance and consensus in peer review. *Science, 214*, 881–886.

Cook, T. D., & Shadish, W. R. (1986). Evaluation: The worldly science. *Annual Review of Psychology, 37*, 193–232.

Cozzens, S. E. et al. (1989). *Literature-based data in research evaluation: A manager's guide to bibliometrics*. Report to the National Science Foundation, Washington, DC.

Cozzens, S. E. et al. (Practitioners' Working Group on Research Evaluation). (1994). Evaluation of fundamental research programs: A review of the issues. Prepared at the request of the Office of Science and Technology Policy. Available from the author.

Gibbons, M., & Johnston, R. (1974). The roles of science in technological innovation. *Research Policy, 3*, 220–242.

Gore, A. (1993). *From red tape to results: Creating a government that works better and costs less*. Report of the National Performance Review, Washington, DC.

Groszyck, W. (1994). Assessment of FY 1994 GPRA pilot project performance plans. OMB memorandum.

Hagstrom, W. O. (1965). *The scientific community*. New York: Basic Books.

Hagstrom, W. O. (1971). Inputs, outputs, and the prestige of university science departments. *Sociology of Education*, 44, 375–397.

Healey, P., Rothman, H., & Hoch, P. K. (1986). An experiment in science mapping for research planning. *Research Policy, 15*, 233–251.

Kostoff, R. N. (1988). Evaluation of proposed and existing accelerated research programs by the Office of Naval Research. *IEEE Transactions in Engineering Management, 35*, 4.

Kostoff, R. N. (1994). Research impact assessment: Where are we now? Unpublished manuscript, Office of Naval Research.

Luukkonen, T., & Stahle, B. (1993). Evaluation of research fields: Scientists' views. *Nord, 15*, Nordic Council of Ministers, Copenhagen.

Martin, B. R., & Skea, J. E. F. (1992). *Academic research performance indicators: An assessment of the possibilities.* Brighton: Science Policy Research Unit, U.K.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations.* Chicago: University of Chicago Press.

The Mitre Corporation. (1978). Evaluative study of the Materials Research Laboratory Program. MTR 7764.

National Institutes of Health. (1994). *Report of the External Advisory Committee of the Director's Advisory Committee on the Intramural Research Program.* Bethedsa, MD: NIH.

National Science Foundation. (1987). *Sources of financial support for research prize winners.* (Report no. NSF 87-87). Washington, DC: Author.

Narin, F. et al. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*, Project No. 704R. Prepared in Fulfillment of Contract NSF C-627 with the National Science Foundation. Cherry Hill, NJ: Computer Horizons.

Narin, F., Olivastro, D., & Stevens, K. (1994). Bibliometrics/theory, practice and problems. *Evaluation Review, 18*(1), 65–76.

National Research Council. (1995). *Evolving the high performance computing and communications initiative to support the nation's information infrastructure.* Washington, DC: National Academy Press.

O'Brien, C. (1994). Quantity no longer counts in Britain. *Science, 264(24 June)*, 1840.

Ormala, E. (1989). Nordic experiences of the evaluation of technical research and development. *Research Policy, 18,* 313–342.

Panetta, L. E. (1994). Memorandum for the Heads of Departments and Agencies designated at Pilot Projects under P.L. 103-162. Washington, DC: Executive Office of the President, Office of Management and Budget.

Phillimore, A. J. (1989). University research performance indicators in practice: The University Grants Committee's evaluation of British universities, 1985–86. *Research Policy, 18,* 255–271.

Research Corporation. (1982). Study of patents resulting from NSF chemistry program. Final report on NSF Contract EVL-8107270, New York.

Rigter, H. (1986). Evaluation of performance of health research in the Netherlands. *Research Policy, 15,* 33–48.

Rist, R. C. (Ed.) (1990). *Program evaluation and the management of government: Patterns and prospects across eight nations.* New Brunswick, NJ: Transaction Publishers.

Roth, W. V., Jr. (1995). Improving government performance. Speech given at the Brookings Institution.

Shadish, W. R. Jr., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice.* Newbury Park, CA: Sage.

Skoie, H. (1993). *EC research and technology policies: Some characteristics of its development and future perspectives.* Oslo, Norway: Institute for Studies in Research and Higher Education.

Small, H., & Garfield, E. (1985). The geography of science: Disciplinary and national mappings. *Journal of Information Science, 11,* 147–159.

Storer, N. (1966) *The social system of science.* New York: Holt, Rinehart & Winston.

Travis, G. D. L., & Collins, H. M. (1991). New light on old boys: Cognitive and institutional particularism in the peer review system. *Science, Technology, & Human Values, 16*(3), 322–341.

U.S. Senate. (1993). Government Performance and Results Act of 1993. 103nd Congress, 1st Session, Report 103-58. Washington, DC: U.S. G.P.O.

Van Raan, A. F. J. (1993). Advanced bibliometric methods to assess research performance and scientific development: Basic principles and recent practical applications. Review report, University of Leiden report CWTS-93-05.

Virgo, J. A. (1977). A statistical procedure for evaluating the importance of scientific papers. *The Library Quarterly, 47*(4), 415–430.

Wiklund, A. (1994). Swedish Natural Sciences Research Council. Interview with Author.

Wye, C. G. (1992). *Evaluation in the federal government: Changes, trends, and opportunities.* San Francisco, CA: Jossey-Bass.

Ziman, J. (1994). *Prometheus bound: Science in a dynamic steady state.* Cambridge: Cambridge University Press.