Forum contributions present essays, opinions, and professional judgments. Forum articles speak to and about the philosophical, ethical, and practical dilemmas of our profession. By design the "Forum" is open to diverse views, in the hope that such diversity will enhance professional dialogue. Standard citations and reference lists should be used to acknowledge and identify earlier contributions and viewpoints. Manuscripts should typically not exceed 15 double-spaced typewritten pages in length unless the paper is invited by the Editor.

Performance Measurement Redux

IRWIN FELLER

ABSTRACT

Recent developments in the United States in the use of performance measurement in science policy and higher education are used to comment further on the Perrin–Bernstein–Winston debate about the effective use and misuse of performance measurement. Particular attention is given to the influence of political/organizational factors and the production processes of agencies on how performance measures are constructed and used. The analysis points to further limitations in the use of performance measurement. In both cases, long-gestating, probabilistic linkages between outputs and outcomes limit the usefulness of mainstream indicators as a measure of current agency performance and as a guide to major, discontinuous resource allocation decisions. Conspicuously absent from many performance measurement undertakings are provisions for evaluating the impact of the undertakings themselves. An updated account of the status of the Government Performance and Results Act (GPRA) indicates that the Act has not had the impacts predicted for it.

"Once more unto the breach . . . " William Shakespeare, "Henry V"

INTRODUCTION

The *American Journal of Evaluations* in 1998 and 1999 contained a debate on the effective use and misuse of performance indicators. The debate was initiated by Perrin's (1998) eight-point

Irwin Feller • Senior Visiting Scientist, American Association for the Advancement of Science and Professor Emeritus of Economics, Pennsylvania State University, 2217 Earth & Engineering Sciences, University Park, PA 16802, USA; Tel: (1) 814-865-0691; E-mail: iqf@psu.edu.

American Journal of Evaluation, Vol. 23, No. 4, 2002, pp. 435–452. All rights of reproduction in any form reserved.
ISSN: 1098-2140 © 2002 by American Evaluation Association. Published by Elsevier Science Inc. All rights reserved.

critique that examined what he termed some basic flaws and inherent limitations in the use of performance indicators for performance measurement. Perrin's eight points included: (1) varying interpretations of the "same" terms and concepts; (2) goal displacement; (3) use of meaningless and irrelevant measures; (4) confounding of cost savings versus cost shifting; (5) obscuring of critical subgroup differences by misleading aggregate indicators; (6) limitations of objective-based approaches to evaluation; (7) uselessness of performance indicators for decision-making and resource allocation; and (8) the inconsistency between a narrow focus on measurement and larger new public management precepts of decentralization and delegation of authority.

Bernstein's (1999) article contained a point-by-point rebuttal. His overarching thesis, however, was that Perrin's catalogue of problems derived not from flaws in the basic concepts behind performance measurement, but instead from "poorly implemented systems that focused too much on process and (collection for collection's sake), as opposed to appropriate use of appropriate measures" (1999, p. 86). Bernstein further suggested that Perrin's recitation of past failures of similar previous performance measurement systems, such as program planning and budgeting systems, was rendered nugatory by the 1993 enactment in the United States of the Government Performance and Results Act (GPRA). GPRA, Bernstein asserted, had the potential to provide for a more systematic integration of planning, budgeting, and performance measurement, and for otherwise avoiding the unreasonably high expectations and pitfalls that plagued earlier initiatives.

A more agnostic position about the use of performance measurement was offered by Winston (1999). Based on experiences in Australia and New Zealand with the use of and debate about performance measurement, he contended that discussions about performance measurement were often confused with those about performance indicators, indicating that Perrin's critique likewise lumped the two concepts and practices together. His experiences, however, also led to a concluding observation that "performance measurement systems need to be assessed across a range of programs and settings, to determine which factors are likely to (a) facilitate the achievement of expected results; (b) lead to unintended outcomes; and (c) act as barriers to effective implementation" (p. 98).

This paper addresses this debate anew, in effect taking up the task proposed by Winston. It does so because the issues raised in this exchange continue to pervade debates in many policy fields, journals, or forums, wherever and whenever the topic of performance measurement comes up. The paper is written from the analytical perspective of an empirically oriented economist who has been involved in the application of performance measurement to science and technology policy and to the research and graduate degree activities of research universities.

However, rather than attempt yet another exegesis on the eight-point critique offered by Perrin, the paper focuses on two themes noted by Winston but which analysis and experience suggest require more detailed and explicit attention. These are the (1) production characteristics of the organization/agency whose performance is being assessed and (2) the political and organizational conditions under which performance measurement systems are adopted and implemented.

The paper adds one further element to the earlier *AJE* exchange. Because GPRA features prominently in Bernstein's stand that Perrin's critiques are dated, this paper capitalizes on the passage of time since 1998–1999 to present additional information about GPRA's impacts and status. By way of overview, this update suggests that GPRA has had limited substantive impacts, and that in the main it is beset with the same limitations as earlier endeavors, and indeed for many of the same reasons stated by Perrin.

Like the observations of most commentators on performance measurement, the paper is heavily conditioned by experience, by seeing how legitimate aspirations for accountability and improved decision-making in fact can and do become dissipated, diverted, and distorted. It thus draws on my various roles as researcher, consultant, reviewer, advisor, commentator, and workshop organizer regarding several aspects of performance measurement and program evaluation. These experiences date back to the early 1970s effort by the Commonwealth of Pennsylvania to implement program–planning–budgeting into the state's budget processes; encompass several benefit-cost and program evaluation studies; include recent studies of the role of strategic planning in research universities; and involvement as a research administrator in the application of performance measurement in an academic institution. These experiences continue to the present in ongoing projects to develop performance metrics for science and technology programs for federal and state government agencies in both the United States and other countries. They also include considerable time and energy recently invested in making a go of GPRA for federal science and technology agencies.

I have chosen two fields—science policy and higher education—to explicate the paper's main propositions. This choice relates to my own experiences, but there are also larger analytical reasons. One is that whatever may be the *general* case for performance measurement as a means of demonstrating accountability, improving performance, and so on, the distribution of performance measurements impacts (whether positive, negative, or non-existent) among agencies does indeed appear to be depend heavily on the *specific political or organizational setting* in which it is applied. Context, in short, counts. Thus, before either prescribing performance measurement as a generically salutary approach or condemning it as the latest public management chimera, it is necessary to examine the characteristics and behaviors of (different types) of organizations under different types of monitoring and incentive systems—the economist's much-examined principal-agent problem (Pratt & Zeckhauser, 1985). Another reason is that these two fields make manifest the proposition that, holding political and organizational contexts constant, the construction of performance measures that are reliable and valid cost depends in large part on the production characteristics of the agency whose performance is being measured.

Let me set these two propositions about the importance of context and production processes into the earlier *AJE* exchange. These propositions suggest, first, that the debate about use and misuse of performance measurement is, in part, a debate about whether the flaws and limitations noted by Perrin are systemic. That is, are the problems of performance measurement tied in some predictable manner to the characteristics of the settings in which they are applied, or are they idiosyncratic outcomes that reflect failures of specific individuals, organizations, or modalities. Second, and more importantly, these two propositions represent a fundamental recasting of Bernstein's statement about appropriate indicators and appropriate use of indicators. For here his propositions are presented as hypotheses, or rather contingent statements, not as givens (or readily obtainable states). What indeed are the appropriate measures to be used in performance measurement (for a given agency/organization)? What, indeed, are the conditions that assure that these measures would be used appropriately?

The first question relates to the state of knowledge, the second to organizational intentions and competencies. In point of fact, the choice of measures to use and assurance that those in power can and will appropriately use the measures are the two primary issues that beset use of performance measurement at the federal level. These same issues also flow beneath the surface in the less visible debates about the use (and impact) of performance measures within research universities.

The following sections treat the propositions about production processes and political and organizational contexts, respectively. They begin with general observations and then turn to the specific effects of performance measurement in the two selected fields. In the account of science policy, this includes a review of GPRA experiences through 2002. Prefacing the debate are three givens. First, performance measurement derives from and is a response to increasing demands upon public sector and not-for-profit organizations to demonstrate accountability to external sponsors and other stakeholders. Second, improved information about organizational performance, both over time and in comparison with like entities has potential to contribute to improved organizational performance (and, in a more recently touted justification, provide a readily comprehensible way in which an organization can communicate its goals and achievements to external publics) (see Behn, 1995; Gormley & Weimer, 1999; Hatry, 1989; Wholey & Hatry, 1992). Third, whatever the specific flaws and limitations of prevailing approaches and requirements may be, performance measurement of some form or another is here to stay. Thus, this paper, like those that form the earlier debate, concludes with recommendations on how to do performance measurement better or—based on a sober distillation of past experiences—at least on how to approach it in a way that even if dedicated and well-intentioned efforts fail to provide indicators and assessment methodologies that significantly enhance accountability, decision-making, or organization performance, the potential for dysfunctional impacts is reduced.

PERFORMANCE MEASUREMENT AND AGENCY PRODUCTION PROCESSES

Wilson has noted that from a managerial view, "agencies differ in two main respects: Can the activities of their operators be observed? Can the results of those activities be observed?" (Wilson, 1989, p. 158). The first factor relates to the measurement of outputs, the second to the measurement of outcomes. In a stylized manner, agencies may have (easily) observable outputs and outcomes; observable outputs but not (easily) observable outcomes; difficult to observe outputs but observable outcomes; and neither observable outputs nor outcomes. Wilson labels the first type of agency a production organization, the second a procedural organization, the third a craft organization, and the fourth a coping organization (Wilson, 1989, p. 159).

These variations in organizational production characteristics are recognized widely in the public performance measurement literature. However, their importance in considering the effective use and misuse of performance measurement systems is too often overlooked or muted. Much of the case for, and indeed early examples cited of, the constructive aspects of performance measurement relate to organizations with easily observable outputs and outcomes—typically, service delivery by production agencies, such as the conditions of city street and parks in state and local governments. Science agencies and universities, by way of contrast, have multiple goals, loosely specified production processes, and probabilistic, long-gestating, and loosely coupled linkages between outputs and outcomes. Essentially, they are mixtures of elements drawn from procedural, craft, and coping organizations, with at most a modicum of elements related to production organizations.

Where recognized, the problematic applicability of a performance measurement approach to agencies or organizations with such production characteristics tends to occur in the initial phases of discussion and debate about the introduction of a performance measurement system. The complexities are frequently subsequently ignored, however, as performance measurement is implemented, as data are aggregated across units, as organizational pressures for completing

performance reports to meet (annual) budget or reporting requirements take hold, and as staff and administrative apparatchiks take on responsibility for compiling and interpreting performance data.

Divergence between the laudatory aspirations held up and held out for performance measurement, on the one hand, and the burdensome, irrelevant and dysfunctional aspects of the impacts of performance measurement as implemented, on the other, are evident in the recent application of performance measurement to science policy and higher education. For science policy, perhaps few economists or other analysts of science policy would go as far as Greenburg, who has characterized accountability rules for research as "endless but vain efforts—akin to trying to capture and weigh a fog—to quantify value from government spending on research" (Greenburg, 2001, p. 4). Many, however, even as they try to develop valid, reliable, and programmatically useful indicators, have noted that attempting to assess the outputs and outcomes of basic research within the framework of existing performance measurement systems, including GPRA, carries considerable risk of distorting and diminishing the public sector's investment (Cozzens, 1997; David, 1995). As often noted, basic research is characterized by lengthy, uncertain production processes, considerable variation across fields of science in the characteristics of these processes, and lengthy gestation periods between and among inputs, outputs, and outcomes.

Similarly, the application of performance measurement to higher education is beset with conceptual, measurement, and operational pitfalls. Birnbaum has described the push for performance measurement in higher education as being "driven by the belief that clarifying goals and measuring progress was the royal road to accountability and efficiency" (Birnbaum, 2000, p. 81). Birnbaum further argues that because higher education's goals are often not clear, it is not possible have a set of performance indicators that can adequately measure, evaluate and reward progress towards them. Expressing a perspective similar to Perrin's, Birnbaum describes the operationalization of performance measurement in universities as involving the selection of performance indicators based on the availability of selected forms of data rather than because the data reflect something of importance.

POLITICAL AND ORGANIZATIONAL ASPECTS OF PERFORMANCE MEASUREMENT

Performance measurement systems and related performance indicators are political instruments used within organizational settings. As such, their impacts represent far more and extend well beyond the standard rationale for them as means by which external stakeholders and organizational leaders monitor an organization's performance, determine its level of financial support, and make evidence-based decisions about which of its activities should be expanded, contracted, reformed, or terminated.

Political Considerations

Performance measurement systems impact on the distribution of authority and influence within an organization, as well as on the forms of evidence deemed legitimate in forming decisions. In my experience, they have had special impact in this regard in universities, less so at the federal level. In the case of the federal government, initial expectations that performance measurement, as reflected in GPRA, would establish a formal set of shared responsibilities

between the executive and legislative branches for shaping agency plans, reviewing performance and determining budget priorities, have not been met. Instead, the constitutionally based, historically shaped division of authorities and responsibilities of the two branches, coupled with fragmented decision-making within the legislative branch has meant that for neither better nor worse, the executive and legislative branches still contend with one another in assessing agency performance and setting budget priorities much as before.

Within universities, however, adoption of performance measurement accords well with Perrin's description of the approach as being rooted in a "top-down hierarchical 'control' model" (Perrin, 1998, p. 375). Voicing a similar assessment, Birnbaum has described performance measurement as part of what he has termed the "second academic management revolution" that occurred in the period 1960–2000. Within universities, the effect of performance measurement has been to substitute quantifiable, objective measures of academic performance for faculty judgments, such as substitution of bibliometrics for collegial reading of a colleague's work. Performance measurement, nested within strategic planning, has, in Rhoade's words, come to reflect "the narrowly and conventionally defined goals targeted by central managers and deans" (2000, p. 58). Relatedly, performance measurement has contributed to the bureaucratization of decision-making, and to a relative elevation of the influence of administrative underlings at the expense of a decrease in the influence of faculty peers and departmental units.

Another important difference in the impacts of performance measurement between the federal government and university setting relates to the existence of checks and balances. Checks and balances are institutionalized safety valves to highlight and possibly correct egregious error or opportunistic behavior—such as the use of performance measurement to achieve goals other than enhanced accountability or improved performance. In the federal setting, the constitutional division of powers between the executive and legislative branches, the division of the legislature into two bodies, and the existence of political parties provide a fertile setting for open debates about the nuts and bolts and use of performance assessments. Examination of any component of a performance measurement undertaking, such as the selection of performance criteria, indicators, methodologies, or generation of data, can be undertaken or underwritten by different branches of government and different parties. The result is the potential for competing assessments, and thus for open debate.

No such system of checks and balances exists within universities. Although typically presented as a grass-roots, bottom-up process in which departments and comparable academic units are given leeway to determine relevant indicators and data sources, universities use of performance measurement tends to be centralized and closed. Few opportunities exist to challenge the accuracy of data or of related analysis, especially as unit measures are aggregated for presentation to university-level decision makers. In this more closed setting, data can be suppressed, misrepresented, or manipulated, with few opportunities for correction or challenge.

The potential for these inappropriate uses of data relates to yet another aspect of the political content of performance measurement, namely, the presence or absence of evaluation. One might expect, or with hopefulness both prescribe and project that major decisions based on performance measurement systems would be systematically evaluated (Wholey, 2001). Thus, tying contemporary management principles of continuous quality improvement and the learning organization together with performance measurement, feedback should exist between and among performance measurement/decision-making/evaluation/performance measurement, and so on.

Such is not the case, however; evaluation is typically the missing link. The demand for evaluation within an organization of its major decisions is spotty, and the quality of internally

performed evaluations likely even more so. The reasons are quite straightforward. Major decisions are investments not only of institutional resources, but also of political capital by senior institutional officials. There is indeed truth and experience behind the adage that victory has many fathers, while defeat is an orphan. Thus, officials of public agencies and university leaders are apt to extol the gains from the panoply of strategic planning/benchmarking/performance measurement approaches when outcomes are positive (though with little effort made to introduce an evaluation design that would allow for testing of alternative explanations), and to ignore or mask setbacks when they occur. More is unlikely in the way of evaluation other than in the case of major fiascoes or changes in administrators.

Moreover, in the case of universities, and, unlike again the federal government, there are no congressional oversight committees or General Accounting Office that can provide independent assessments of performance or review the quality of institutional performance claims. The setting, as indicated by my own experiences and other contemporary research (Rhoades, 2000), facilitates irrelevant, unreliable, and potentially biased assessments of the actual performance of universities and an uncritical, unevaluated assessment of the relationship of performance measurement to actual performance.

Organizational Considerations

The level of organizational competencies also influences the extent to which performance measures are used well and wisely or poorly or misleadingly. Use of performance indicators is not a simple matter, with technical issues of performance measurement inextricably being linked with how appropriately measures are used. Even the most widely used and seemingly straightforward measures are laden with technical complexities. Without attempting to delve deeply into these complexities, recent examples from the use of bibliometrics and patent statistics highlight the pitfalls that await even the most knowledgeable users.

Bibliometrics are a widely used indicator of the performance of scientists and their organizations. Over time, citation counts of increasingly refined detail—who is cited how many times by whom and in what journals—have displaced publication counts as the primary output measure, with citation "impact factors" increasingly becoming the quantitative proxy for quality. Contributing to the use of these measures has been the increased availability of readily accessible general and customized databases, such as the Science Citation Index. Obtaining such counts is now a relatively straightforward if not costless matter.

The construction and interpretation of citation data are beset with recognized complexities. For example, researchers' propensity to cite articles, as well as conventions for the ordering of authors, varies among disciplines. In the social sciences, prevailing practice is to list contributors in the order of their relative contribution to the article, with graduate research assistants and technicians usually listed towards the tail end of the list of names or acknowledged in footnotes. In science and engineering fields, it is more customary, but not universal, for the graduate student who performed the actual research experiment upon which the manuscript is based to be listed first, followed by other contributors, with the senior author—the individual in whose laboratory the research was done—to be listed last.

Two linked sources of errors of measurement are possible here. One is that without detailed knowledge of authorship conventions by field, or even of the distribution of practices within a field, the identification of the individual/institution/country who is to be credited with "performance" may be misreported. The second is that some citation services truncate the list of authors; in the case of articles with multiple contributors, particularly when the convention

is to list the senior researcher last, this practice may delete the participation and contribution of key actors (National Academy of Sciences, 2000).

Another potential error is misinterpretation of data. In a highly visible and well-regarded study, Narin, Hamilton, and Olivastro (1997) used citation linkages between U.S. patents and scientific research papers to demonstrate a strong and increasing linkage between what they termed "public science," that is, research papers whose authors acknowledged external support from U.S. government agencies, and U.S.-invented patents. The article thus buttressed the case for federal government support of basic research and has been frequently referred to in this context.

The technical issue here is the source and thus interpretation of "non-patent references" in patent citations. A review of prior knowledge, patented or otherwise, is a standard part of the patent application process and is intended to ensure the novelty of the invention and to identify the limits of its claims. These references, which are on the title page of patents, come from two sources: applicants and examiners. The perspective and behaviors of the two parties can vary, however. This variation is one of several criticisms raised by Meyer (2000), who contended that many of the citations counted by Narin et al. were inserted by examiners, not inventors, and therefore cannot be used to demonstrate that the inventor drew upon the publicly funded research.

The details of this debate extend beyond the scope of this paper. What is relevant is its highly technical nature, which one may safely assume is likely to be beyond the ken of most performance measurement units or staffers in most organizations. Thus, it is not simply that the wrong things may be measured and that choice of measures leads to goal displacement, but also that those doing and using the measures do not always understand what they mean.

PERFORMANCE MEASUREMENT, SCIENCE POLICY, AND GPRA

Whatever its flaws, GPRA is not monolithic. The Act recognizes that not all agencies can readily produce objective, quantifiable, and measurable statements of goals and accomplishments. It thus permits an agency, in consultation with and approval by the Office of Management and Budget (OMB), to develop an alternative form of performance measurement and reporting that employs descriptive statements to assess whether a program has achieved minimal levels of effectiveness or is successful. Among federal science agencies, only the National Science Foundation (NSF) has applied for and received permission to use the alternative approach. Overall, though, the National Academy of Sciences has judged that federal science agencies have made "good-faith" efforts to develop reporting procedures that comply with the requirements of GPRA (National Academy of Sciences, 2001, p. 3). Still, science agencies continue to struggle with reconciling GPRA's provisions with the production characteristics of basic research. The following paragraphs sketch out the sources of some of these struggles.

(a) The uncertain what, when, how, where, and why of the impacts of scientific discoveries are staple observations in the history of science and technology. Game theory, for example, for which John Nash received a Nobel Prize in Economics, has transformed research and education in economics, and its impacts have spilled over to change the way problems are defined and approached in other social and life sciences. Yet few accurately predicted its impact at the time that Nash did his seminal work, which was 1949–1950. As noted by Selten, who in 1994 shared the Nobel Prize with Nash

and Harsanyi, "Nobody would have foretold the great impact of the Nash equilibrium on economics and social science in general. It was even less expected that Nash's equilibrium concept would ever have any significance for biological theory" (as quoted in Nasar, 1988, p. 98). Nor is this a case of a rose born to bloom unseen. Rather, in the case of game theory, there was a flurry of activity around its use, followed by disillusionment with its explanatory power (Nasar, 1988, p. 122).

Any assessment of the returns to public investment in game theory, say in 1960, would have given a negative value; an assessment taken in 1980 would have been far more positive; and that taken today would give it the venture capitalist's star. In brief, the 5-year planning horizon established by GPRA is too brief a time to adequately assess the outcomes of science. Indeed, for several types of research characterized by "batch process" relationships among construction of research facilities, data collection, and analysis and interpretation of data, such as high energy physics or longitudinal health or social science studies, it is limited even as a means of assessing conventional research output, such as publications.

Put differently, were performance measurement of the type represented by GPRA truly to take hold, the result would be pressure on agencies and performers (added to that already said by many to exist as a result of the discipline-based peer review system) to engage in risk-averse behavior—the frontiers of knowledge would be located just outside one's backyard, well interior to the horizon. Concern that the performance measurement will deteriorate to this state, with indicators required of so many major discoveries or significant outcomes per reporting period, is palpable among federal agency science managers.

(b) Whatever the degree of exactitude and credence attached to measures of scientific output, the transformation of these outputs into outcomes is frequently so complex, indirect, and subject to factors and forces beyond those of the funding agency or researcher as to vitiate the meaningfulness of the output measures (David, Mowery, & Steinmueller, 1992). Consider the case, say, of research funded by National Institute of Health (NIH, 2002) that leads to the detection of a genetic marker for detecting a disease, which, given early detection, is treatable. The socially desired outcome of reduced morbidity and mortality depends, at a minimum: on the rate, extent, and manner in which this publicly funded research is converted into a commercially available test kit; successful approval of this test kit by regulatory approval processes; awareness by the individuals susceptible to the disease that a test exists followed by action; awareness and incorporation into practice of the test kit by doctors; and the individual's access to health care services, which is a function of the economics of the nation's health care delivery system. What is an appropriate performance indicator: the rate of scientific discovery, or the rate by which the adverse health effects of the disease are reduced? Moreover, in terms of annual or even 5-year budget decisions, what if the changes in these two kinds of indicators varies by some appreciable amount, or even for a time, given the long-term, uncertain linkage between the two, diverge significantly? What if an alternative test or discovery is made in the interim which makes the first product obsolete before its in general use?

Agencies have a bifurcated but logically consistent response to these questions. They eschew responsibility for those outcomes about which they have limited control (or for which performance measures may raise questions about agency impacts), pointing at all times to the cumulative, past benefits from the portfolio of their activities. They also

claim credit for societally beneficial outcomes—the perennial anecdotes or nuggets that stud agency performance reports—even though their contributions may have been only one part in a larger, multiparty, multisector story.

(c) Perhaps the most telling limitation of performance measurement as applied to science policy is that whatever its value may be in tracking past performance and monitoring dimensions of current activities (Guston, 2000), it is of limited value for prospective decisions. Past performance obviously is an important contributor to the shaping of decisions for the future. It may indeed be an appropriate heuristic if the future is presumed to resemble the past—why else require resumes on research proposals or job applicants? But the past is of less help in gauging major scientific turning points and decisions involving truly expensive, discrete, singular scientific endeavors. How, for example, should the Congress or Administration react to a proposal from the high energy physics community for the U.S. to be centrally involved, possibly taking the lead, in the construction of the Next Linear Collider? This device promises to be "one of the great scientific adventures of our time," but has an estimated price tag of between $5 and 7 billion, with the host country paying about two thirds of the bill and Germany contemplating a competitive initiative (Seife, 2002).

The quality of an agency or research group's past performance may be of some help in reaching a decision, especially in inducing little faith in units that have been shown not to be performing well. But as voiced by national science leaders, the decision to go ahead with such an enterprise entails many factors that are not readily converted into quantitative terms, such as the excitement of pursuing the unknown and national leadership.

(d) Little evidence exists to indicate that GPRA has any demonstrable impact on the pattern of federal support for science and technology, whether in the aggregate or distribution among federal agencies. Likewise, as noted both by external observers and OMB, there is little to indicate that GPRA has affected the core decisions of federal science agencies with respect to the size and priorities of their budgets, modes of research support, or internal organizational arrangements (National Academy of Sciences, 2001, p. 5). Likewise, Radin, in a review of GPRA written from a public management perspective, has concluded that, "Viewed as a whole, GPRA has failed to significantly influence substantive policy and budgetary processes. Instead, its use of administrative rhetoric has caused it to collide with institutional, functional, and policy/political constraints that are a part of the American decision-making system" (Radin, 2000, p. 133). In the main, it appears that GPRA reporting processes have been essentially partitioned off from operational decisions and units. It has become a staff function, involving input from and (considerable!) effort on behalf of operating units and external advisory bodies but essentially prepared by headquarters staffs.

The recent doubling of NIH's budget over a 5-year period has reflected broad-based, bipartisan support for fundamental research, the influence of legislators identified with these agencies, and the strong support of industry, which continues to capitalize on the public sector's investment in the foundational research from which commercially profitable new products and processes may be developed. The budgetary success of NIH has now led NSF's congressional and other supporters to propose a similar 5-year doubling of its budget, on the grounds that balance needs to be re-established among supporting fields of science. The doubling mantra has now spread to other agencies, such as The Department of Defense (DoD). Not all federal

agency budgets for science and technology however, are projected for increases. The National Institute of Standards and Technology's (NIST) applied technology and manufacturing extension programs, the Department of Energy's (DE) energy conservation programs, and the Environmental Protection Agency's (EPA) environmentally related research and development have confronted recurrent efforts first by a Republican-controlled Congress in the mid-1990s and more recently by the Bush Administration to terminate or reduce their funding. These efforts are manifestly tied to ideological stances, not to the ability of these agencies to "document" performance results to any lesser degree than NIH, NSF, or DoD.

Indeed, although GPRA continues to have many advocates within the federal government, the public management community, and a growing performance measurement industry, recent published and oral statements by federal agency officials and researchers, including several initially supportive of GPRA, point to growing disappointment. The Bush Administration, for example, clearly intends to bring a "results" orientation to management of the federal government, including tying budget decisions to past performance. But, in contrast to Bernstein's expectations about the legislation, the Administration sees little value to GPRA in reaching its objective. As voiced by Administration representatives in public forums, its view is that GPRA has become a paper exercise, burdening federal agencies with extensive reporting requirements but having little impact on agency behaviors, and most importantly offering little in the way of integration between performance review and budget priorities. The President's Management Agenda report, issued through OMB, for example, in referring to GPRA states that:

> After 8 years of experience, progress toward the use of performance information for program management has been discouraging. According to a General Accounting Office survey of federal managers, agencies may, in fact, be losing ground in their efforts to building organizational cultures that support a focus on results (2002, p. 27).

To augment GPRA, the Bush Administration, on May 30, 2002, in a joint memorandum prepared by the Office of Science and Technology and the OMB, promulgated a set of investment priorities and investment criteria for research and development that were to be used in developing FY2004 budget requests. The memorandum set forth three key criteria: relevance, quality, and performance. Relevance relates to clearly stated plans that the proposed R&D investments relate to national priorities, specific presidential priorities, agency missions, and relevant fields. Quality relates to the mechanisms for awarding R&D grants and contracts—with competitive, merit-based processes singled out as the preferred mode—and to periodic assessment of current and past R&D efforts. Performance relates to the management of an agency's R&D programs in a manner that produces indentifiable results.

The memorandum presents a nuanced statement of the difficulties of applying the criteria to basic research ("the Administration is aware that predicting and assessing the outcomes of basic research in particular is never easy. Serendipitous results are often the most interesting and ultimately may have the most value"). Still, it contends that, "there is no inherent conflict between these facts and a call for clearer information about program goals and performance towards achieving these goals."

The R&D Investment Criteria memo was followed by OMB's issuance on July 12, 2002, of a 36-page set of instructions for a new Performance Rating Assessment Tool (PART), to be used by all agencies. PART is a "diagnostic tool that relies on objective data to inform evidence-based judgments to assess and evaluate programs across a wide range of issues related to performance." Moreover, although at places presented as complementary to GPRA, the PART instructions also inform agencies that although they may use GPRA performance

measures as a starting point, they may also have to revise these measures significantly (to reflect a focus on outcomes), and even delete unnecessary measures. Thus, rather than leading to a unified, consistent approach to planning and budgeting, recent federal government initiatives in performance measurement have led to a bifurcated, potentially competitive environment. Well aware of the powers of both OMB and the Congress over their appropriations and activities, federal science agencies (as indeed all agencies) must of necessity respond to each branch's set of criteria and reporting formats, even where they are different.

The OMB R&D investment criteria and the implementation of these criteria via the PART process are too new at the time this is written to permit comment on their impacts. OMB's openness to review and critique during the formulation of the criteria, including National Academy of Sciences and other expert workshops, points to considerable awareness of the uses and misuses of this approach, and some hope for judicious and constructive application. However, much the same could be said about the early phases of similar earlier performance measurement initiatives. Only after sometime will we see what transpires between an agency head and his/her cognizant OMB examiner in balancing the competing demands for annual performance indicators and the reasoned and well-recognized claims that basic science activities be exempted from them. Here, past performance, necessarily if unhappily, provides little grounds for optimism.

None of these developments should be a surprise, for they derive from defining characteristics of public policy formulation in the U.S. As described by Radin, three of these characteristics are, "the structures of fragmented decision-making in the U.S., the imperatives of several decision-making functions (particularly the differences between budgeting, management and planning), and the dynamics of politics and policy making in the American political system" (Radin, 2002, p. 111).

To note here only the fragmented political character of federal government decision-making, and then only its most obvious structural characteristics, federal appropriations entail a complex process, requiring at a minimum agreement between the preferences and priorities of the executive branch and those of the Congress (National Academy of Sciences, 1995a, pp. 62–69). Within Congress, responsibility for an agency's budget and operations are apportioned among authorizations, appropriations, and government oversight committees. GPRA is often seen as the legislative offspring of the Senate Committee on Governmental Affairs. This committee's responsibilities include the organization and reorganization of the executive branch of the government, but do not extend to the purse strings. In the case of federal science policy (and indeed across the broad swath of other functional areas), there is little to indicate that this committee's concerns or assessments of agency performance relate to the actions of the cognate Senate committees (and subcommittees!) that oversee NSF, NIH, DE, NASA, or DoD's authorizations or appropriations or indeed of the Senate at large.

PERFORMANCE MEASUREMENT IN HIGHER EDUCATION

Beyond its general flaws, performance measurement in higher education serves as a conservative influence, preserving existing organizational arrangements rather than enhancing institutional performance. As used on some campuses, it has had especially deleterious effects on (a) intellectual curiosity, (b) interdisciplinary research, and (c) reform of graduate education. These effects have occurred as a result of both intended and unintended consequences.

(a) A bibliometric cottage industry has developed in several academic disciplines, with economics being a prime example. Among their several features, these studies measure the degree to which articles in a selected set of journals are cited by authors publishing in other highly cited journals. They thus serve to document (and reinforce) existing perceptions of a hierarchy of journals. In this respect, they can serve as a useful signal to faculty about where their work is likely to have greatest visibility and impact. But there is a manifest intellectual downside to this development as institutions assign increased weight to these citation measures in strategic and resource allocation decisions regarding both individual and academic units.

Citation patterns among "fields" may be characterized in terms of quantity and reciprocity. In the case of the social sciences, asymmetry is a marked feature, with economists, for example, seldom citing articles outside their discipline, even as researchers in other fields cite economics articles (Pieters & Baumgartner, 2002). Coupled with use of only a limited number of "core" discipline-based journals to evaluate individual or unit performance, this asymmetry means, to use the language of economics, that an economist who publishes outside of the list of journals used in the institution's performance measurement system or who is cited by researchers who publish in these outside journals—say *Science* or *AJE*—generates zero personal marginal product (as computed for promotion or salary increases) and zero marginal organizational output (Stephan, 1996).

The consequences of performance measurement in such a setting is towards a narrowing of intellectual horizons. Faculty come to eschew questions, methodologies, and dissemination to other than their disciplinary peers. Either because they are risk-averse (even if tenured) or because of pressures to stay within disciplinary bounds placed on them by department heads, who are reacting, in turn, to larger organizational incentive positive and negative, their choice of research questions is constrained. The sign posts of the frontiers of knowledge are placed at the boundaries of their department's backyard.

(b) Calls are widespread for a greater interdisciplinary orientation in the research and graduate education of U.S. higher education. The calls arise from many sources: from the dynamics of science, which leads to the formation of new "fields" or disciplines (e.g., cognitive sciences, nanoscience); from professional societies and employers who have noted that addressing both basic and applied problems of society and industry requires knowledge and skills that transcend a single academic discipline; and from federal agencies that, both in response to and in anticipation of the above trends, have adopted funding mechanisms that support interdisciplinary research centers and graduate degree programs (e.g., NSF's Engineering Research Centers and Integrative Graduate Education and Research Traineeship Program; Brainard, 2002). Fitting these programs into the traditional disciplinary-based department and college structures of universities is a long-standing issue (Feller, 2002; Lattuca, 2002). The problems have been magnified by rote adherence in academic strategic planning undertakings to the precept, "if you can't measure it, you can't manage it." An unrecognized corollary of that dictum was that if something hasn't been measured, it doesn't exist. Since by definition reporting and budget arrangements for interdisciplinary graduate degree and research programs tend to fall outside of conventional reporting units, their salient achievements—research awards, quality and placement of students—were easily overlooked. As noted by the Government–University–Industry Research Roundtable, for

example, "Interdisciplinary programs are 'orphans' within the fiscal bureaucracy of the university. These programs are at a further disadvantage since most of the university's planning efforts are based on the fiscal structure. Thus, interdisciplinary programs play less prominent role in the long-range planning of the university" (Government–University–Industry Research Roundtable, 1994, p. 7). Rhoade's more recent study of strategic planning experiences in 40 departments at four research universities likewise points to negative impacts of the combination of strategic planning and performance measurement on interdisciplinary programs (Rhoades, 2001). (On two fronts, this situation may be changing. First, increased coverage of degree fields and attention to interdisciplinary studies is projected for the next NRC study. Second, interdisciplinarity features prominently in the strategic plans, circa 2000, of may research-intensive universities.)

(c) An unintended but deleterious consequence of performance measurement on interdisciplinary graduate education, a variant of the above problem of the absence of a measurement category leading to diminution of the value of an activity, flows from the structure of the 1995 National Research Council's (NRC, 1995) report, *Research-Doctorate Programs in the United States*. The report provides an assessment of the "quality" of graduate programs in 41 "fields," organized, in the main, about long-standing discipline-based definitions, with some allowance for the rise (since the prior 1982 assessment) of some fields in the broad area of biological sciences.

As indicated by interviews at a cross-section of universities conducted as part of a series of studies on the competitive structure of the U.S. research university system (Feller, 1996, 2000), a dominant feature of academic strategic planning in the post-1995 feature has been for institutions to seek to "advance" in the NRC quality rankings. This objective, reinforced by the ubiquitous hold of the strategic planning maxims of "selectivity" and of "not being all things to all people," has led many institutions to concentrate only on fields already listed in the NRC report. Because they did not enter into the calculations by which colleges or universities improved on quality or reputational measures, non-listed programs, including interdisciplinary programs, tended to be given short shrift or eliminated. However, the absence of listings, in the NRC study related to limited resources and to the conceptual and empirical difficulties of measuring interdisciplinary programs, not necessarily to a qualitative assessment of the importance of omitted areas of science.

Strikingly, another major source of performance data on the status of graduate education similarly is blind to the existence of interdisciplinary programs. The 2002 NSF-NIH Annual Survey of Students and post-doctorates in science and engineering (S&E) asked respondents to list the name of their S&E department or program. The instructions for answering this question were as follows: "A student should be reported in only one department. Students enrolled in interdisciplinary/institutional programs should be counted only once, by their 'home' department and institution." These instructions not only hide the quantity of interdisciplinary graduate education occurring on U.S. campuses, but also serve as a disincentive for academic units to engage in such programs because they cannot document (claim 'credit' for) their activities to higher level college or university performance objectives.

The consequences of inadequate measures and reification of selected measures over understanding and vision are not trivial. They reinforce the status quo in discipline-based organizations of graduate education, in the face of national trends for more than one-half of new Ph.D.s to find work in non-academic settings and of calls from national research and

professional associations for a broadening and reshaping of U.S. graduate education (National Academy of Sciences, 1995b). And in a not insignificant manner, the use of performance measurement by universities may contribute to the obstacles confronting the development of graduate degree programs in evaluation, particularly those purposely adopt an interdisciplinary perspective, seeking both on the input (faculty) side and on the output side (policy area and placement of graduate students) a greater breadth than currently exists (Stufflebeam, 2001).

CONCLUSIONS AND RECOMMENDATIONS

The problem with the current use of performance measurement, clearly in the cases of science policy and higher education and, relatedly also, it seems likely in similar fields, is not the "inherent flaws" with the measures—measures can and should be improved. Here I think, as Winston has suggested, Perrin's (1998) choice of wording is unhelpful in lumping together issues of logic, technique, and implementation. Rather, the above amalgam of general and specific features suggest that the value of performance measurement in fostering accountability, contributing to improved organization performance, and communicating an organization's goals and results is limited by: (1) the imperfect state of knowledge about what these measures should be, how to construct them, and the administrative feasibility and cost-effectiveness of data collection and analysis, and (2) political and organizational contexts. These contexts variously tend to transform a limited but reasonable technique that meets legitimate demands for accountability, effectiveness, and efficiency into a substantively vacuous but effort-demanding undertaking. Performance measurement can be and has been a form of symbolic politics that provides political coverage for an organization with few significant impacts on the organizations. In my experience, it has also been a vehicle and veneer for opportunistic behavior directed at capturing resources and control but with dysfunctional impacts on the organization's long-term competitive capabilities.

The potential for misguided, mistaken, and malevolent use of performance measurement within larger organizational structures leads to several recommendations. None of them are novel, but the above accounts of the use of performance measurement in science policy and higher education indicate that they need to be taken more seriously in considering the design, adoption, and implementation of performance measurement systems. The above accounts also indicate that appropriate selection and use of indicators are only the first steps in promoting enhanced organizational performance. Monitoring and evaluating the impacts of performance measurement, as Perrin and Winston noted, are also essential components of a systemic approach to improved organizational performance.

(a) The first recommendation is, Do no harm. This is not a platitude. To adopt this recommendation is to understand that the potential for misuse of performance measurement is systemic, not idiosyncratic. Do no harm is a frequently voiced expression of senior federal agency officials and program managers, pleadingly uttered in the hope of warding off the several dysfunctional consequences of GPRA and its ilk (including, the new OMB criteria and PART procedures). Their concern, grounded in experience, is that the informed, nuanced understanding expressed in open forums by senior organizational officials about the limits or complexities of applying performance measurement systems or specific indicators can quickly deteriorate to mechanistic, rigid demands

by junior examiners, committee staffers, or academic apparatchik's for specific but specious or irrelevant annual metrics.

(b) Agreement must be established among sponsors, users, and performers about organizational objectives, and consequently about the specification and measurement of outputs and outcomes. This means, in part, that agreement is reached among the parties involved in conducting the performance assessment and operating the program about the relevant and correct program theory(ies) for linking inputs, outputs and outcomes before measures are selected.

(c) Closely related to (b) is the need, in Perrin's words, to actively involve stakeholders in "developing, reviewing, and revising measures . . . " and to actively involve them "in interpreting findings and identifying implications" (Perrin, 1998, p. 376). The recommendation here is conceived in a more tactical manner than indicated by the breadth of Perrin's statement, or by Newcomer's account of the increased involvement of citizens in setting performance objectives and measures for public and non-profit programs or the use of participatory evaluation within the evaluation profession (Newcomer, 2001). Rather, it is addressed specifically to Perrin's closing clause, which involved interpreting findings and identifying implications. Where outputs and/or outcomes are difficult to measure or to relate to one another, as in the case of science policy and higher education, channels and forums for correction and redress must be established within a performance measurement system to avoid errors and misinterpretations, accidental or intentional. As applied to universities, the recommendation implies that the centralization and bureaucratization of analysis and interpretation of data that has accompanied adoption of performance measurement by universities must be lessened if performance measurement is to be used correctly and wisely.

(d) Care should be taken in accepting the claims made by agencies or evaluator that gains in performance have been accurately recorded and that these gains are causally related to changes associated with the appropriate use of the appropriate measures. Realism, coupled with continued hope and effort, are required here. Experience suggests that Perrin's admonition, and indeed that of many other members of the evaluation profession, that evaluation be considered an integral, indispensable component of performance measure is a chord that, while not always falling on tone-deaf listeners, is at least muffled in many settings by many competing and louder sounds. But the admonition must be stated, and repeated as often as necessary. Continuing efforts need to be made to highlight to public sector officials that evaluation is an essential element in a comprehensive effort to improve organizational performance through adoption of a performance measurement system.

(e) The rhetoric on behalf of performance measurement must change. It is not enough to justify performance measurement systems on Churchillian grounds, that it is less bad than any alternative mechanisms. That is not so. Other approaches, including the hallowed if now seemingly out-of-fashion mode of expert panels, need reconsideration. Of course, no alternative is free of problems of measurement, comparability, selection bias (both in membership composition and data sources), cost, and more. Nevertheless, alternative performance assessment frameworks offer the prospect of at least as robust a reliance on the quantitative orientation typified by performance measurement programs, and also of "equal" opportunity for consideration of other forms of evidence, as well as opportunity for experience and insight. The NAS review of agency performance under GPRA, for example, argues that although the Act "strongly encourages

agencies to evaluate their programs annually through the use of quantitative measures so that progress can be followed with clear numerical indicators," in its view, "research programs, especially those supporting basic research, cannot be meaningfully evaluated this way annually" (National Academy of Sciences, 2001, p. 9). Instead, the panel recommends that these programs be "evaluated over a somewhat longer term through expert review, which has a long tradition of effectiveness and objectivity" (National Academy of Sciences, 2001, p. 39).

CODA

Economists, by self-selection or by training or both, are dedicated to the pursuit of efficiency. Empirically oriented economists in particular invest considerable intellectual effort in using theory to develop valid measures, in collecting data, and in analyzing and interpreting these data. Therefore, economists should, and indeed do, express a natural predilection to endorse performance measurement as a means of enhancing the performance of public sector and not-for-profit organizations (for which, in their analytical framework, conventional market metrics are not readily available). But experience and observation also count. Performance measurement, at least in its current state of technical development, legislative requirements, and actual implementation, may be a necessary medicine for many agencies and organizations, but its use needs to be surrounded by bold-face cautions about potential harmful side effects.

REFERENCES

Behn, R. (1995). Here comes performance assessment—and it might even be good for you. In A. Teich, et al. (Eds.), *AAAS science and technology policy yearbook* (pp. 257–264). Washington, DC: American Association for the Advancement of Science.

Bernstein, D. (1999). Comments on Perrin's effective use and misuse of performance measurement. *American Journal of Evaluation*, *20*, 85–93.

Birnbaum, R. (2000). *Management fads in higher education*. San Francisco, CA: Jossey-Bass.

Brainard, J. (2002). U.S. agencies look to interdisciplinary research. *Chronicle of Higher Education*, June 14.

Cozzens, S. (1997). The knowledge pool: Measurement challenges in evaluating fundamental research programs. *Evaluation and Program Planning*, *20*, 77–89.

David, P. (1995). Difficulties in assessing research and development programs. In A. Teich, et al. (Eds.), *AAAS science and technology policy yearbook* (pp. 293–301). Washington, DC: American Association for the Advancement of Science.

David, P., Mowery, D., & Steinmueller, E. (1992). Analyzing the economic payoffs from basic research. *Economics of Innovation and New Technology*, *2*, 73–90.

Feller, I. (1996). The determinants of research competitiveness among universities. In A. Teich (Ed.), *Competitiveness in academic research*. Washington, DC: American Association for the Advancement of Science.

Feller, I. (2000). Social contracts and the impact of matching fund requirements on American research universities. *Educational Evaluation and Policy Analysis*, *22*, 83–89.

Feller, I. (2002). *New organizations, old cultures*. Paper presented at the American Association for the Advancement of Sciences 2002 Meeting, Boston, MA.

Gormley, W., & Weimer, D. (1999). *Organizational report cards*. Cambridge, MA: Harvard University Press.

Government–University–Industry Research Roundtable. (1994). *Stresses on research and education at colleges and universities: Institutional and sponsoring agency responses*. Washington, DC: Author.
Greenburg, D. (2001). *Science, money and politics*. Chicago, IL: University of Chicago Press.
Guston, D. (2000). *Between politics and science*. New Brunswick, NJ: Rutgers University Press.
Hatry, H. (1989). Determining the effectiveness of government services. In J. Perry (Ed.), *Handbook of public administration* (pp. 469–482). San Francisco, CA: Jossey-Bass.
Lattuca, L. (2002). *Creating interdisciplinarity*. Nashville, TN: Vanderbilt University Press.
Meyer, M. (2000). Does science push technology? Patents citing scientific literature. *Research Policy, 29*, 409–434.
Narin, F., Hamilton, K., & Olivastro, D. (1997). The increasing linkages between U.S. technology and public science. *Research Policy, 26*, 317–330.
Nasar, S. (1988). *A beautiful mind*. New York: Simon & Schuster.
National Academy of Sciences. (1995a). *Allocating federal funds for science and technology*. Washington, DC: National Academy Press.
National Academy of Sciences. (1995b). *Reshaping the graduate education of scientists and engineers*. Washington, DC: National Academy Press.
National Academy of Sciences. (2000). *Experiments in international benchmarking of U.S. research fields*. Washington, DC: National Academy Press.
National Academy of Sciences. (2001). *Implementing the Government Performance and Results Act for research*. Washington, DC: National Academy Press.
National Institutes of Health. (2002). *Final FY2003 GPRA performance plan. Revised final FY2002 GPRA annual performance plan, and FY2002 GPRA annual performance report*. Washington, DC: U.S. Department of Health and Human Services, Appendix 1.
National Research Council. (1995). *Research-doctorate programs in the United States*. Washington, DC: National Academy Press.
Newcomer, K. (2001). Tracking and probing program performance: Fruitful path or blind alley for evaluation professionals. *American Journal of Evaluation, 22*, 337–341.
Perrin, B. (1998). Effective use and misuse of performance measurement. *American Journal of Evaluation, 19*, 367–379.
Pieters, R., & Baumgartner, H. (2002). Who talks to whom? Intra- and interdisciplinary communication of economics journals. *Journal of Economic Literature, 40*, 483–509.
Pratt, J., & Zeckhauser, R. (1985). Principals and agents: An overview. In J. Pratt & R. Zeckhauser (Eds.), *Principals and agents* (pp. 1–35). Cambridge, MA: Harvard Business School Press.
Radin, B. (2002). The Government Performance and Results Act and the tradition of federal management reform: Square pegs in round holes? *Journal of Public Administration Research and Theory, 10*, 111–135.
Rhoades, G. (2000). Who's doing it right? Strategic activity in public research universities. *The Review of Higher Education, 24*, 41–66.
Seife, C. (2002). Report backs collider and an expanded field. *Science, 295*, 783ff.
Stephan, P. (1996). The economics of science. *Journal of Economic Literature, 34*, 1199–1235.
Stufflebeam, D. (2001). Interdisciplinary Ph.D. programming in evaluation. *American Journal of Evaluation, 22*, 445–455.
Wholey, J. (2001). Managing for results: Roles for evaluators in a new management era. *American Journal of Evaluation, 22*, 343–347.
Wholey, J., & Hatry, H. (1992). The case for performance monitoring. *Public Administration Review, 52*, 604–610.
Wilson, J. (1989). *Bureaucracy*. New York: Basic Books.
Winston, J. (1999). Performance indicators-promises unmet: A response to Perrin. *American Journal of Evaluation, 20*, 95–99.