Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 103–114
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

The unintended consequences of metrics in technology evaluation☆

Ronald N. Kostoff[a,*], Elie Geisler[b]

[a] *Office of Naval Research, 875 N. Randolph Street, Arlington, VA 22217, United States*
[b] *Stuart School of Business, Illinois Institute of Technology, 565 W. Adams Street, Chicago, IL 60661, United States*

Received 1 June 2006; received in revised form 4 January 2007; accepted 8 February 2007

Abstract

This paper describes science and technology (S&T) metrics, especially impact of metrics on strategic management. The main messages to be conveyed from this paper are: (1) metrics play many roles in supporting management of the S&T enterprise; (2) metrics can influence S&T development incentives; (3) incorrect selection and implementation of metrics can have negative unintended consequences on the research and research documentation generated and (4) before implementing metrics, an organization should identify and evaluate the intended and unintended consequences of the specific metrics' implementation, and identify the impact of these consequences on the organization's core mission.

Keywords: Metrics; Research evaluation; Scientometrics; Research policy; Research documentation; Science and technology; Strategic management; Unintended consequences

1. Introduction

The outcomes from science and technology (S&T) underpin modern economies and defense capabilities. Government and industry provide the bulk of resources for S&T development, with government supplying the majority of basic science resources and industry contributing substantial resources to more advanced technology development. In both sectors, S&T accountability procedures have become more requested, more visible, more frequent, and more formal. Questions persist about the most credible methods for insuring accountability to satisfy a variety of stakeholders.

Peer review, the expert judgment by specialists within a given discipline, has been the traditional method used for S&T accountability. Performance metrics (the counting of research activity, outputs, impacts, and quantification of outcomes) tend to be advocated by S&T decision makers who may not be technical specialists, but want independent credible measures of S&T quality and progress that could support resource allocation decisions. The consensus of most of the S&T community is that peer review is the preferred approach to be used for S&T accountability

☆ The views in this paper are solely those of the authors and do not represent the views of the US Department of the Navy, any of its components, or the Stuart School of Business.

* Corresponding author.
E-mail address: kostofr@onr.navy.mil (R.N. Kostoff).

1751-1577/$ – see front matter
doi:10.1016/j.joi.2007.02.002

(evaluation/assessment), strongly supported by the use of appropriate metrics. However, the selection of appropriate metrics remains an outstanding issue. This paper aims to provide some insight to the role of metrics in the S&T accountability process, and the criteria for selection of metrics most appropriate to the problems being addressed. In particular, because S&T metrics can serve as S&T development incentives, this paper highlights the positive and negative intended and unintended consequences for S&T that could result from incorrect selection of S&T metrics.

The remainder of the paper describes:

- S&T accountability.
- Effects of S&T expenditures:
  - structure,
  - flow.
- Attributes of S&T metrics:
  - qualitative/quantitative metrics,
  - prospective/retrospective metrics.
- Impact of metrics selection on strategic management.
- Unintended consequences from metrics selection.
- Re-balancing quantitative and qualitative metrics.

2. S&T accountability

What is S&T accountability, how is it performed, and how does it relate to metrics?

The S&T organization can be viewed from a decision-consequences perspective as having two major components: (1) expenditures of S&T funds and (2) the S&T-related effects resulting from those expenditures. S&T accountability is the identification and assessment/evaluation of the S&T-related effects resulting from the S&T expenditures. S&T accountability is performed through evaluations/assessments of the expenditures and resulting effects by a combination of (1) experts in the relevant S&T disciplines and (2) experts in technical strategic objectives and mission areas impacted by the S&T under evaluation. Metrics are the instruments that enable the identification and assessment/evaluation of the S&T-related effects. The challenge is to identify the suite of metrics instruments that will enable credible accountability without being overly burdensome, unwieldy, or expensive.

Accountability is achieved through the demands of legal requirements and organizational or managerial needs. Much of what is reported in the effort to assess and evaluate S&T is open to some degree of interpretation as to the nature and content of the evaluation, and the degree to which such content has fulfilled the demands for accountability.

Metrics of S&T and its outcomes play a very significant role in contributing to the perception by stakeholders of the S&T organizations that their reporting indeed results in accountability. Different constituencies and stakeholders have differing standards and benchmarks of what constitutes acceptable assessment and what can be described as acceptable accountability. Thus, S&T metrics would have to effectively measure the S&T expenditures and its outcomes in such a manner that the differing perceptions would be satisfied (Rubenstein & Geisler, 1988).

Similarly, the various stakeholders of the S&T organization utilize the S&T evaluation results in different ways and for different reasons and objectives. This means that these stakeholders may wish to incorporate the metrics of the assessment/evaluation into their own systems of reporting and managerial control (Loch & Staffan-Tapper, 2002). This action by the stakeholders places additional burden on the S&T organization when it strives to achieve accountability, particularly in the selection of metrics for the assessment of S&T.

Evaluation for the purpose of accountability addresses both the expenditures for S&T and the S&T effects. The expenditures are evaluated primarily by the use of accounting methods, and are subject to a plethora of rules and regulations imposed by various government bodies for fiscal and monetary objectives of transparency and honesty. Organizational managers are also interested in the accurate and comprehensive accounting of S&T expenditures.

But, it is in the category of S&T effects from such expenditures that the main difficulties arise in establishing adequate evaluation with metrics that will satisfy the various constituencies.

3. Effects of S&T expenditures

The effects of S&T expenditures can be classified into two major categories: (1) structure and (2) flow. The structure represents characteristic features of the S&T being conducted (e.g., merit, approach, team, risk, status), while the flow can be conceptualized as the flux of products from the S&T being conducted (e.g., activity, outputs, impacts, outcomes). The challenge mentioned above translates into selecting metrics as evaluation criteria that will effectively evaluate the nature of the structural and flow effects of the S&T expenditures.

3.1. Structure

The structure category is a set of features of the S&T activity itself, resulting from the S&T expenditures. How many, and what types of, evaluation criteria are required to provide adequate insight to the structural characteristics of the project/program being reviewed? Large numbers of criteria become unwieldy operationally, provide excessive resolution, and mask/dilute the major insights and findings from the review. Very small numbers of criteria provide inadequate insight/resolution to the project's/program's structure.

The evaluation of the structure category is done for two basic purposes. The first is to allow senior and program/project management the tools and the ability to identify and understand existing structural problems. This diagnostic function of the evaluation effort provides managers with the capability to intervene, correct problems, and to learn from their experience.

The second purpose is to enable managers to link the structure category to the organization's strategic objectives. This link allows managers to assess the value of the project/program to the organization's mission, goals, and strategic milestones. By doing so managers are also able to create a common terminology with external stakeholders who are now apprised of the relevance of the S&T effort to the survival, growth, and competitiveness of the organization (Godener & Soderquist, 2004; Kerssens-Van Drongelen & Bilderbeek, 1999).

A minimum set of evaluation criteria for the structure category that balances adequate insight/resolution with operational flexibility consists of the following five criteria: *merit*, *approach*, *team quality*, *risk*, and *status*. These criteria are differentiated chronologically by the S&T development cycle stage in which they first exert influence on the decision-making process (planning → portfolio selection → review → transition) as follows.

Merit addresses the importance of the S&T being reviewed to both the larger S&T community and the S&T organization's mission, specifically, whether the appropriate overall objectives (in the context of the organization's strategic objective) are being pursued by the project/program under review. The focus is on S&T and strategic goals, not on approach. *Merit* exerts influence on the decision-making process starting at the earliest stages of S&T planning, and continues to exert influence on the portfolio selection, review, and transition stages. Examples of merit metrics could include research merit and relevance to the organization.

Approach addresses the conduct of the S&T project/program, specifically whether the conduct will lead to attainment of the specified S&T project/program goals and objectives. *Approach* exerts influence on the decision-making process at the portfolio selection stage, and continues to exert influence on the review and transition stages. Examples of approach metrics could include balance between experiment and theory, balance between resources and objectives, state-of-the-art of instrumentation, and coordination with other organizations.

Team Quality addresses the competence of the people who manage and perform the S&T. *Team Quality* exerts influence on the decision-making process at the portfolio selection stage, and continues to exert influence on the review and transition stages. Examples of team quality metrics could include past publication quality of the team, citations, awards, and honors.

Risk addresses the degree to which the S&T project/program will achieve its stated goals and objectives, and has some relation to the quality of S&T performers and the approach selected. *Risk* exerts influence on the decision-making process at the portfolio selection stage, and continues to exert influence on the review and transition stages. Examples of risk metrics could include probability of achieving S&T objectives, probability of impacting long-range objectives, and the probability of successful demonstration.

Status addresses the progress that has been made on the S&T development, and has some relation to the quality of S&T performers and approach, and to risk. *Status* exerts influence on the decision-making process at the review stage, and continues to exert influence on the transition stages. Examples of status metrics could include technology readiness level, objectives completed, and technical milestones completed.

3.2. Flow

The flow category contains the S&T product-related effects resulting from the S&T expenditures. These product-related effects can be classified into four categories (*activity*, *output*, *impact*, *outcome*), differentiated by their temporal distance from the time the S&T funds were expended.

Activity reflects the S&T infrastructure generated from the initial S&T expenditures. It starts immediately after the portfolio selection stage, and continues through all successive stages. The *activity* is under direct control of the S&T resources recipient. Examples of activity metrics could include numbers and types/quality of people conducting the S&T, numbers and types/quality of equipment used in the S&T, and numbers and types/quality of facilities used for the S&T. There is some overlap between the team quality criterion used for structure evaluation in the review and transition stages, and the people quality component of activity.

The *activity* category has also been described as the process by which S&T is actually generated with the resources that the expenditures allow the organization to procure. The metrics of this category measure such resources. They also provide the basis for the assessment of how well such resources have been processed and applied in the generation of S&T (Council on Competitiveness, 2004; Lessenius, Missinen, Raujiaven, & Sulonen, 1998).

Output reflects the initial products from the S&T under review. It starts well after the portfolio selection stage, continues through the review and transition stages, and may continue even after transition due to long lag times. The *output* is under direct control of the S&T resources recipient. Examples of output metrics could include numbers and quality of journal papers, numbers and quality of patents, and numbers and quality of professional presentations.

The output category refers to the set of proximal or immediate outputs from S&T (Geisler, 2000; Rubenstein & Geisler, 1988). These are the results from S&T which include primarily bibliometric measures of written outputs and patents generated by the S&T activity (Weingart, 2005).

In the decision process of S&T, these immediate outputs are measured throughout the evaluation stages (from portfolio selection to transition), but they represent the first crop of *measurable* results from S&T which are of relevance to those who conduct the S&T activity and to the other stakeholders in S&T.

Impact reflects the influence of the S&T under review on the external S&T and potential user communities. It typically starts years after the initiation of *output*, and can continue years after transition (decades in some cases). The *impact* is not under the control of the S&T resources recipient, but rather under external control, typically (but not exclusively) by other members of the S&T community. Examples of impact metrics could include numbers and quality of paper and patent citations, numbers and quality of awards/honors, and numbers and quality of downstream development plans altered due to S&T outputs.

The impact category includes those results from S&T which may be described as “intermediate” outputs (Geisler, 2000). These are results that impact S&T stakeholders beyond the focal S&T producer. These results influence selected activities, plans, decisions, and the further applications of the outputs that S&T has generated. Such influence typically occurs when downstream entities apply, integrate, or implement S&T outputs into their own operations.

Outcome reflects the far downstream effects of the S&T under review on the larger scale goals of the economy and society. It starts well after transition, perhaps even decades afterwards, and can continue for many years or decades. The *outcome* is not under the control of the S&T resources recipient, but rather is impacted by changing user interests, environmental, political, financial, legal, international, and other non-technical considerations. Examples of outcome metrics could include lives saved, cost savings, increased performance and capability, improved rate of return, and improved quality of life.

These effects are also described as ultimate outputs and they measure significant social and economic results that occur when downstream S&T stakeholders apply the outputs from S&T. For example, business enterprises may implement S&T outputs (devices or techniques), thus gaining increased productivity, cost-savings, and cost-cutting. Similarly, social entities gain from the application of S&T results, thus engendering outcomes that have consequences in the areas of health, education, transportation, and other venues of social importance.

Activity, *output*, and *impact* are (mathematically) products of quantity times unit quality. Thus, if publication outputs are being evaluated, not only are the numbers of publications important, but the quality of each publication is important as well. These three flow criteria can be separated mathematically into their quantity and quality components for simple estimations, but for credible S&T evaluation, the quantity–quality product is required. Since *outcomes* tend to be fewer than the above three flow quantities and temporally very distant, but larger in magnitude of effect, separation of *outcomes* into quantity and quality components is not useful. Detailed analyses by experts are required for credible evaluation of *outcome* results.

The categories of structure and flow effects (potential evaluation criteria categories) resulting from S&T expenditures have been defined, and some metrics examples have been provided. The question now arises as to the intrinsic properties of these metrics, and how these properties affect operational use of the metrics.

4. S&T evaluation and metrics: a review

The search for metrics of S&T has been an area of investigation for several decades. Werner and Souder (1997) reviewed the literature and concluded that S&T metrics are a function of such criteria as users' needs, type of S&T, the available data, and the resources at the disposal of the evaluators.

Using a single metric to evaluate S&T (in toto or by stages of its development) has been largely discounted in the literature. Such a macro-measure lacks the appropriate theory that would justify its selection, and lacks the necessary agreement on which metric to choose. Thus, a unique bibliometric may be a measure of S&T output (or *immediate* outputs) but cannot be a sufficiently comprehensive metric of the entire output of S&T (Geisler, 2000; Weingart, 2005).

The resulting view calls for a broad set of metrics describing the various stages and attributes of the S&T continuum. The issue therefore is the selection of a set of metrics that will sufficiently measure the S&T phenomenon. The initial criteria for selection of metrics include at least the available pool of metrics, the objectives of the evaluation effort, and the preferences of the evaluators and the organizations in the S&T continuum.

Geisler (2000) reviewed the literature and proposed seven categories or groups of metrics in the overall available pool: (a) investments in S&T; (b) economic-financial; (c) commercial and business-related; (d) bibliometrics; (e) patents; (f) peer review and (g) organizational, strategic, and managerial metrics. These categories may include over 100 different metrics. What are the basic criteria that allow the evaluators to make the first cut from this pool to extract a preliminary candidate set of metrics?

Geisler (2000) suggested six such criteria: (1) *methodology* (what metrics should do and how), such as: quality of the data provided by the metrics, and ease of data collection, manipulation, and interpretation; (2) *ontology* (characteristics of the construction of metrics) such as: validity and relative convergence with other metrics; (3) *organization* (what metrics should accomplish) such as: relevance to organizational objectives and cost-effectiveness; (4) *availability* (are the metrics available in the pool from which the evaluators extract the metrics); (5) *accessibility* (can the metrics be accessed, even when available) and (6) *affordability* (are the metrics – when combined in a meaningful set – within the economic capabilities of the evaluation effort).

Each of the seven categories of metrics described above has strengths and weaknesses. The literature suggests that a balanced S&T evaluation should have metrics representing each category, because these categories have metrics that measure different aspects and stages of the S&T structure and flow. Some of these metrics are quantifiable while others can only be obtained in qualitative terms.

5. Attributes of metrics for S&T evaluation

Once a preliminary selection of S&T metrics is conducted per the categories listed above, there emerges the need to introduce another level of classification of these metrics. This level considers the overall attributes of the metrics selected for the evaluation pool of potential metrics. The classification by these attributes allows for the establishment of initial benchmarks for metrics, which will in turn help the evaluation to apply the most effective S&T metrics at their disposal.

S&T metrics have two fundamental intrinsic characteristics that span the 'objectivity-time' continuum. The 'objectivity' characteristic ranges from objective (quantitative, machine-supplied data) to subjective (qualitative, human-supplied data). The temporal characteristic ranges from retrospective (looking backward in time) to prospective (looking forward in time). Each of these two intrinsic characteristics will be discussed in more detail below.

5.1. Qualitative/quantitative metrics

The two fundamental approaches to S&T evaluation, peer review and performance metrics, use two intrinsically different types of metrics. Peer review generally uses *qualitative* (subjective) metrics, and performance metrics uses largely *quantitative* (objective) metrics. Both types of evaluation also use metrics that are a hybrid of qualitative and quantitative. Purely qualitative metrics use data supplied by humans. These subjective types of data are typically judgments of items (e.g., manuscript quality, level of project risk, degree of project innovation, level of project technological readiness, and quality of researchers). Purely quantitative metrics use data supplied by machine or systems embedded within the organization with minimal human assumptions. These objective types of data are typically counts of items (e.g., numbers of papers, numbers of patents, numbers of transitions, numbers of researchers, and revenues generated).

Even quantitative-objective data need to be analyzed, put in context, and assembled for meaningful comparisons and analysis. This is done by the evaluators following a given set of principles, standards, and benchmarks. By themselves, quantitative data are devoid of relevance to the phenomenon under study—unless they are purposely inserted into an evaluation scheme by means of analytical frameworks (Kostoff, 2006; Mann, Mimno, & McCallum, 2006).

Hybrid metrics use data supplied by machine, supplemented by substantial human judgment on which machine data is to be selected for analysis and how the machine data is aggregated to quantify the metric. The people who perform the data selection and aggregation to quantify the hybrid metrics require substantial knowledge of the underlying S&T, and perhaps business, marketing, and application data as well, depending on the specific hybrid metrics selected. This is contrasted with the simple counting of papers, patents, and citations used for the purely quantitative metrics, where many assumptions or much judgment are not required from the analyst, nor is any understanding about the underlying S&T required by the analyst. These objective/subjective hybrid metrics are typically outcome-related (cost-benefit ratios, rate of return, cost savings, or their national security/medical equivalents).

5.2. Some attributes of the metrics

The subjective qualitative metrics applied to S&T evaluation today tend to have the following characteristics:

- More complex in concept than simple item counts.
- More expensive to obtain.
- More manually intensive, and less amenable to automation.
- More training required for implementation and interpretation.
- Less consistency across projects.

These qualitative metrics require not only the expertise of human evaluators, but also their relative objectivity for credible interpretation. Hence, evaluators who have strong interests in the organization being evaluated or who are employed by the S&T organization may have strong views and opinions that tend to distort their interpretation of qualitative metrics.

However, they also usually possess extensive knowledge about the S&T organization, its objectives, and its operations. They can be relatively more amenable to training as evaluators and have more knowledge about the organization's culture and internal processes. To assume a satisfactory level of effective evaluation and analysis of qualitative metrics, both internal and external resources should be procured.

The objective quantitative metrics used in S&T evaluation today have their origins in industrial-age production measures. Quantitative metrics based on past data tend to involve S&T productivity counts. These types of productivity metrics are (relative to the subjective qualitative metrics):

- Simpler in concept.
- Relatively inexpensive to obtain.
- Easily amenable to automation.
- Implemented and interpreted with minimal training.

Although these metrics have attributes that make them attractive to use, there is a need for analysis by human evaluators. The current state of knowledge in S&T metrics does not allow for an algorithm whereby such quantitative metrics may be manipulated by machines or expert systems and analyzed to produce sensible evaluation results without human intervention (Ethiraj & Levinthal, 2004; Katz & Cothey, 2006).

The criteria categories defined for *structure* (merit, approach, team, risk, status) tend to be qualitative metrics. The criteria categories defined for *flow* (activity output, impact, outcome) tend to be (1) quantitative for the counting component of activity, output, and impact, (2) qualitative for the non-counting components of these criteria categories, and (3) hybrid for the outcomes. For both types of metrics, one important selection consideration today is minimal disruption to the organization's operations (Egghe & Rousseau, 1990).

Both quantitative and qualitative metrics have different levels of certainty and credibility, depending on whether they use past, present, or future data. The relation between time perspective, credibility, and application will now be examined.

5.3. Prospective/retrospective/present metrics utilization

Prospective use of metrics involves prediction/estimation of the metrics' values at future points in time. The uncertainty/credibility associated with the metrics' values increases with the length of prediction/estimation time. As an example, a cost-benefit estimate of market implementation in 2020 of products resulting from S&T performed today would be a prospective hybrid metric, with substantial uncertainty. As another example, the impact on quality of life in 2020 of S&T performed today would be a prospective qualitative metric, also with substantial uncertainty.

Conversely, retrospective use of metrics involves tabulation of the metrics' values from past points in time. Retrospective tabulation is an inherently more certain and credible process. As an example, the cumulative number of citations (from papers accessed by the Science Citation Index) received over the past decade by papers published in the mid-1990s would be a retrospective quantitative metric, with a high degree of certainty. As another example, the impact on quality of life in 2000 of S&T performed in the 1960s would be a retrospective qualitative metric, with relatively reasonable certainty (Klavans & Boyack, 2006).

Finally, present metrics involve specification of the metrics' values at the present time. As an example, the quality of the approach of an ongoing S&T project would be a present qualitative metric. As another example, the specific performance status today of a fighter aircraft prototype under development would be a present quantitative metric.

The rationale for, and value of, using metrics retrospectively, in the present, or prospectively depends on the intended application. Retrospective use of metrics tends to be valuable for:

- Generating lessons learned from past development.
- Marketing based on actual achievements.
- Identifying management environments conducive to successful development.
- Rewarding personnel involved in successful development.
- Accountability based on past performance.
- Regulatory, fiduciary, and other requirements.

However, retrospective use of most quantitative metrics (e.g., number of citations recorded, number of awards received, amount of revenue generated) and qualitative metrics (e.g., quality of demonstrated impact on S&T, quality of awards, quality of life enhancement demonstrated) is of limited value for some S&T management purposes. These include program modifications (directions, budgets, personnel) to correct real-time performance problems, new program selection based on potential impact and payoff, and marketing based on potential payoff.

In particular, the availability of impact or especially outcome data resulting from S&T program execution typically occurs too far downstream from the S&T program initiation to influence future program execution (research direction, budgets, personnel). For example, paper citation data would not be available for credible evaluation purposes until at least six (or preferably more) years after an S&T project had been initiated, given the reality of publication delays for the initial published papers and for the subsequent citing papers. Market implementation data would not be available for one or two decades after S&T project initiation (for most technologies). These long time intervals between S&T program initiation and the availability of data for evaluation purposes preclude the use of this retrospective data to impact the original S&T program's decision-makers or influence the S&T program's direction in a timely manner.

Nevertheless, in special cases, the use of short-term retrospective metrics (e.g., number and quality of papers recently published, number and quality of researchers recently hired) could provide timely data to partially influence program execution decisions. More importantly for this type of application (influence present program execution decisions) would be the use of 'present' metrics from recent peer reviews (e.g., research team quality, research approach quality, progress status, technology readiness distribution and associated quality of distribution bands). Having this current data would help with decisions taken to correct problems (with the S&T project) identified by the evaluation/metrics so that they would be applied to the people, allocations, and budgets associated with those problems.

Prospective use of quantitative and qualitative metrics (e.g., estimated impact on S&T, estimated sales revenue streams, estimated operational cost savings, estimated quality of life enhancement, estimated increase in organizational capabilities) is quite valuable for some of the applications unavailable to retrospective use of metrics, including:

(1) new program selection based on potential impact and payoff, and
(2) marketing based on potential payoff, especially marketing at early stages of the S&T development.

Unfortunately, the data generated prospectively are far more uncertain than the retrospective data. Prospective S&T metrics data should be generated by researchers with a thorough understanding of the S&T at all phases of its proposed evolution trajectory from the present to its future estimation point, if such data are to have credibility.

If selected and applied properly, metrics can be of substantial benefit for strategic management (and marketing) at all stages of the S&T development cycle shown above. But what is the relation between selection of S&T metrics and strategic management of S&T development?

The need for adequate metrics is conducive to a situation whereby the existing pool of specific metrics in effect may dictate the nature of the link. This becomes a case of the "tail wagging the dog," as John Kenneth Galbraith (1985) had suggested that the foreign policy of the United States would be influenced by the range of its military aircraft. This is the case where the ability to measure S&T strongly influences the nature of *measurable* objectives and milestones of the strategy of the organization.

The evaluation of S&T via the use of metrics of structure and flow contributes to the strategy of the organization in at least two complementary manners. First, S&T allows for an enhanced measure of certainty in the development of organizational processes and roadmaps to deal with environmental threats of competition, acquisition of resources, and performance (Loch & Staffan-Tapper, 2002).

Second, S&T provides long-term value to the strategy of the organization by offering alternative paths for growth and survival. This is done via the ability of S&T to assist in the application of core competencies, and by opening up future avenues for the organization.

Yet, these processes require adequate metrics that permit the establishment of benchmarks, contribute to the justification of the selection of certain strategies, and help to provide control and maneuverability of strategic paths, as well as direction and corrective actions.

5.4. S&T metrics and the balanced scorecard

The emergence of the balanced scorecard (Kaplan & Norton, 1996) as an example of a managerial evaluative framework was predicated on the view that a management evaluation system should go beyond financial measures to include a variety of measurable goals and processes. This model has four perspectives: financial, customer, internal business processes, and learning and growth. All these perspectives are designed to influence the vision and strategy of the organization. Each perspective is assessed by objectives, measures, targets, and initiatives.

Metrics are a central component of the balanced scorecard (Kaplan & Norton, 2006). They provide the facts upon which diagnostic, feedback, and corrective actions are instituted within the vision and the strategy of the organization. Metrics are used to measure inputs, processes, and outcomes.

In this vein, S&T metrics are a crucial element of the measurement effort within the balanced scorecard. In at least the "internal business processes" and the "learning and growth" areas of the scorecard, S&T plays a significant role in contributing to the achievement of the vision.

In summary, without adequate metrics of S&T, there is little input into the measurement content of the balanced scorecard, with the result that a facts-based assessment becomes nearly impossible to accomplish. One cannot improve, direct, correct, or evaluate what one cannot adequately measure.

Initially, the strategic plan provides the criteria for the development of S&T (and business) metrics. After a while the S&T metrics are used to assess progress and to decide on future actions. The importance of these metrics is now magnified to an extent that they may dictate modifications in the strategy, far beyond their original role and importance. Once the metrics are incorporated in a system such as the balanced scorecard, their actual measures effectively determine present and future actions by management. S&T metrics have the potential to irrevocably influence the technical direction and well-being of the organization. It is therefore essential to ensure that the selection of these metrics and their measurement is diligently and carefully accomplished.

6. Linking S&T metrics and strategic management

There is an extensive literature advocating the congruence, integration, and complementarity of S&T and strategy (Das & Van de Ven, 2000; Souitaris, 2002). The reasons for this include: S&T is involved in every activity in the value chain; S&T and strategy must follow the same direction to achieve congruence of organizational goals, and the alignment of S&T and strategy allows for better accountability of S&T, with acceptable milestones and with standards for performance which go beyond the limited range of S&T and include those of the entire organization (Couzzarin & Percival, 2006).

The advantages of having S&T and strategy in a position of alignment or even integration of goals and outcomes require a system of metrics. This alignment is predicated upon the capability of the organization to measure the outcomes and impacts of S&T, and to measure the objectives, milestones, and performance of the overall organization (Geisler, 2000; Godener & Soderquist, 2004).

6.1. Impact of S&T metrics selection on strategic management

In many research project/program evaluations, productivity (in the broader context of including all the flow categories defined previously) assumes a central role, and in a very real sense is where the 'rubber meets the road'. Not only are the numbers of activities, outputs, impacts, and outcomes important for determining productivity, but the quality of these productivity items is equally or more important. Unfortunately, there is a severe imbalance today between the use of retrospective quantitative and qualitative indicators in the reporting of S&T productivity. Due to the simplicity and other advantages of obtaining the retrospective activity/output/impact quantitative metrics data, compared to obtaining the qualitative metrics data shown above, much of the S&T productivity reported today is primarily quantitative. This can have many negative unintended consequences. The following sections relate these consequences to the types of metrics used and how the metrics can be selected to both minimize the negative unintended consequences and promote positive intended consequences.

In practice, one strong reason for the selection of the simple retrospective quantitative productivity metrics is to make minimal time demands on the organization S&T program officers and field research performers. To accomplish this arbitrary (but understandable) objective of minimal intrusion on the organization's operations, the data available from ordinary organizational business operations become the major source of data to populate the metrics. The available data from organizational business operations thus serve as the pro forma drivers for determining the metrics to be used, which in turn determines the objectives whose S&T progress will be gauged by the metrics. This is the reverse of what would be desired from strategic management of S&T and has been challenged by systems such as the balanced scorecard. The following steps are advisable:

- Set objectives for desired outputs and outcomes.
- Define metrics that would gauge S&T progress toward meeting these objectives.
- Determine the data required to populate these metrics.

When managers engage in "cherry-picking" metrics that are easily obtained, they weaken the link between S&T and strategy—a link essential to a robust evaluation of the strategic plan and direction of the organization. Within a model such as the balanced scorecard, the various perspectives are effectively "shortchanged" because adequate S&T metrics may not be available.

S&T metrics should provide a trusted and workable representation of the factors and drivers that make up the structure and flow of S&T, and that can be used to align S&T with the strategic goals of the organization. This crucial objective of S&T evaluation is not achieved when S&T is measured by any set of metrics—simply because they are easily attainable, but not scrupulously selected to carry out the task of evaluation.

What are the consequences (to S&T alignment with strategy and for S&T development) of available organizational business data determining the metrics selected for evaluation?

7. Unintended negative consequences from metrics selection

For data gathering in physical, environmental, engineering, and life sciences applications, care is taken to insure that the measuring instruments have minimal impact on the state of the system being measured. Except for the fundamental limitations on measurement precision imposed by Heisenberg's uncertainty principle, which becomes of concern only at very small scales, these instruments are becoming more able to exert minimal influence on states of systems being measured.

For the S&T development cycle, the situation is intrinsically different (Shupe & Behling, 2006). The metrics employed have the potential to influence the S&T development trajectory. Additionally, they have the potential to serve as incentives and thereby distort the development results and objectives severely, sometimes in very unintended directions. In particular, if production-based productivity metrics are perceived by the S&T sponsors and performers to be the dominant form used for S&T evaluation, the incentives for S&T sponsors and performers are likely to:

- alter the types of S&T performed,
- alter the types of S&T documents and outcomes produced, and
- alter the direction of the S&T effort/program.

This may lead to a concerted effort to maximize output quantity. These distorted incentives lead to negative unintended consequences. Weingart (2005), for example, summarized a few of these negative unintended consequences. He suggested that they are likely to:

- Increase publication counts by fragmenting articles.
- Propose conservative but safe research projects. The objective here is to minimize the risk of failure, and insure the continual supply of publications.
- Increase publication quantity at the expense of quality. This is especially true when quality metrics are not included in the measurement suite.
- Increase bias toward short-term performance as opposed to long-term research capacity.
- Increase bias toward conventional approaches.

Perhaps the most serious negative impact of expanded use of conventional production-based 'productivity' metrics in the S&T development cycle would be the strong negative incentives provided for radical discovery and innovation, counter to the recommendations in the National Innovation Initiative Report of the Council on Competitiveness (2004) to strongly promote this type of radical discovery and innovation. As shown in Kostoff (2006), much of truly radical discovery and innovation will involve cross-disciplinary extrapolation of concepts. Unfortunately, very strong negative incentives exist for cross-disciplinary or inter-disciplinary research (Kostoff, 2002). This is primarily because:

- Much time is required for the performers to learn multiple disciplines or new disciplines, leaving less time for publishing, and reducing publication (and patent) outputs.
- Much time is required for coordinating and synchronizing research across disciplines, subtracting time that could be devoted to generating publications and other outputs.
- Journal review of trans-discipline manuscript submission is much more difficult, resulting in higher manuscript rejection rates, and reducing publication outputs.
- Grants are more difficult to obtain because of the trans-disciplinary review problem, reducing metrics based on research support funds obtained.
- All these effects impact tenure and honors/awards negatively, radically reducing the use of metrics based on achievements (Siggelkow & Rivkin, 2006).

In the same vein, an insightful article by Rick Weiss (2005) decried the decline of non-applied curiosity-driven research. He attributed the reason for the decline to 'deliverables' (e.g., specific products, profits, outcomes) becoming the dominant force driving research agendas.

S&T metrics can also be focused on the link between S&T productivity and quantitative measures of strategic objectives. Geisler (2000) identified several key organizational measures such as cost-effectiveness, contributions to organizational sales and similar quantitative indicators of organizational success. Performers of S&T are effectively obligated to conduct S&T whose outcomes will contribute to such quantitative indices. Preference is given in the funding of S&T whose metrics will be conducive to measurable outcomes and benchmarks of the organization (Pun & White, 2005).

Most, if not all, quantitative metrics that are perceived by the research performers to influence research funding and research rewards will, on average, steer the research away from curiosity-driven multi/inter-disciplinary high-risk research characteristic of radical discovery and toward mono-disciplinary low-risk low-discovery research. When meeting quantitative milestone and deliverable targets, or maximizing output quantity, become the de facto research goals, any diversions that cost time/money and involve risk will of necessity be eliminated. These production-based de facto incentives have the potential to drastically change the nature of research performed, especially basic research.

For over a decade American corporations have been closing down corporate S&T laboratories, thus effectively eliminating basic or curiosity-driven research (Geisler, 1995). The impacts of globalization and outsourcing are among the several phenomena that have combined with the effects of production-based metrics to focus on relevancy and to weaken the exploratory nature of S&T.

What can be done to counter these negative incentives for radical discovery and innovation?

8. Re-balancing quantitative and qualitative metrics

To correct the de facto imbalance of quantitative to qualitative retrospective productivity metrics, qualitative metrics should be added to the suite of criteria used for S&T evaluation. These metrics could include (but not be limited to): innovation potential, creativity, discovery potential, originality, level of risk, probability of success, potential for mission or strategic impacts, research merit, research approach quality, potential for transition, program executability, team quality, technology readiness level, exploitation of external S&T, and leveraging of external S&T.

It should be noted that for inclusion of more qualitative metrics in the suite of evaluation instruments/metrics, there is no guarantee that the present desire for minimal disruption of research sponsors and performers during the evaluation process will be achieved. Additionally, for inclusion of either quantitative or qualitative metrics that have been determined starting from objectives and goals rather than available organizational business operations data, there is also no guarantee that the present desire for minimal disruption of research sponsors and performers during the evaluation process will be achieved.

A general rule for metrics selection to insure some minimum balance between quantitative and qualitative productivity metrics is that every purely quantitative productivity metric should be accompanied by one or more qualitative metrics. Thus, if one output measure displayed is number of journal papers produced, the quality of those papers should be added. If an output measure is number of transitions produced, the quality and potential impacts of those transitions should be added. If an output measure is number of researchers developed, the quality of those researchers should be added. The necessity for the performers to now maximize the quantity–quality product, rather than to maximize the quantity only, will lead to different (more desirable) types of research.

9. Conclusions and recommendations

The complexities of selecting, employing and interpreting S&T metrics invariably lead to the creation of intended and unintended consequences to S&T-performing organizations. A partial solution to attenuating negative unintended consequences is the rebalancing of quantitative and qualitative metrics, and the incessant effort to select appropriate metrics.

The selection of appropriate metrics will also involve difficult tradeoffs among: (1) providing positive incentives to meet organizational and even national objectives; (2) generating cost-savings; (3) improving quality due to increased accountability and (4) considering the full cost of implementing S&T metrics.

There is no “magic bullet” that will solve the problems listed in this paper and allow us to generate and implement a suite of S&T metrics without any negative consequences. Vigilance, understanding of the phenomenon, and applying corrective actions are the hallmark of good S&T management.

We therefore recommend that before implementing specific metrics for application to any part of the S&T development cycle, the organization should identify and evaluate the intended and unintended consequences from the implementation of specific metrics. In addition, the organization must identify the impacts of these consequences on the strategic goals and the core mission of the organization, and to appropriately adjust and correct for such impacts.

References

Council on Competitiveness (2004). *National Innovation Initiative Report: Innovate America*. Washington, DC.

Couzzarin, B., & Percival, J. (2006). Complementarities between organizational strategies and innovation. *Economies of Innovation & New Technology*, *15*(3), 195–217.

Das, S., & Van de Ven, A. (2000). Competing with new product technologies: A process model of strategy. *Management Science*, *46*(10), 1300–1316.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation, and information science*. New York: Elsevier Publishing Company.

Ethiraj, S., & Levinthal, D. (2004). Bounded rationality and the search for organizational architecture: An evolutionary perspective on the design of organizations and their evolvability. *Administrative Science Quarterly*, *49*(3), 404–437.

Galbraith, J. (1985). *The anatomy of power*. New York: Houghton-Mifflin Company.

Geisler, E. (1995). When whales are cast ashore: The conversion to relevancy of US basic research and universities. *IEEE Transactions on Engineering Management*, *41*(1), 3–8.

Geisler, E. (2000). *The metrics of science and technology*. Westport, CT: Greenwood.

Godener, A., & Soderquist, K. (2004). Use and impact of performance measurement results in R&D and NPD: An exploratory study. *R&D Management*, *34*(2), 191–213.

Kaplan, R., & Norton, D. (1996). *The balanced scorecard*. Cambridge, MA: Harvard Business School Press.

Kaplan, R., & Norton, D. (2006). *Alignment: Using the balanced scorecard to create corporate synergies*. Cambridge, MA: Harvard Business School Press.

Katz, S., & Cothey, V. (2006). Web indicators for complex innovation systems. *Research Evaluation*, *15*(2), 85–95.

Kerssens-Van Drongelen, & Bilderbeek, J. (1999). R&D performance measurement: More than choosing a set of metrics. *R&D Management*, *29*(1), 35–46.

Klavans, R., & Boyack, K. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, *57*(2), 251–263.

Kostoff, R. N. (2002). Overcoming Specialization. *BioScience*, *52*(10), 937–941.

Kostoff, R. N. (2006). Systematic acceleration of radical discovery and innovation in science and technology. *Technological Forecasting and Social Change*, *73*(8), 923–936.

Lessenius, G., Missinen, M., Raujiaven, K., & Sulonen, R. (1998). The interactive goal panel: A methodology for assigning R&D activities with corporate strategy. In *Proceedings of the International Conference of IEMC*.

Loch, C., & Staffan-Tapper, U. (2002). Implementing a strategy-driven performance measurement system for an applied research group. *Journal of Product Innovation Management*, *19*(3), 185–198.

Mann, G., Mimno, D., & McCallum, A. (2006). Bibliometric impact measures leveraging topic analysis. *JCDL*, June 11–15.

Pun, K., & White, A. (2005). A Performance measurement paradigm for integrating strategy formulation: A review of systems and frameworks. *International Journal of Management Reviews*, *7*(1), 49–71.

Rubenstein, A., & Geisler, E. (1988). The use of indicators and measures of the R&D process in evaluating science and technology programs. In D. Roessner (Ed.), *Government innovation policy* (pp. 185–204). New York: St. Martin's Press.

Shupe, C., & Behling, R. (2006). Developing and implementing a strategy for technology deployment. *Information Management Journal*, *40*(4), 52–57.

Siggelkow, N., & Rivkin, J. (2006). When exploration backfires: Unintended consequences of multilevel organizational search. *Academy of Management Journal*, *49*(4), 779–795.

Souitaris, V. (2002). Technological trajectories as moderators of firm-level determinants of innovation. *Research Policy*, *31*(6), 877–898.

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences. *Scientometrics*, *62*(1), 117–131.

Weiss, R. (2005). Our incredible shrinking curiosity. *The Washington Post*, B-01.

Werner, B., & Souder, W. (1997). Measuring R&D performance: State of the art. *Research Technology Management*, *40*(2), 34–42.