Future Assessment Task Group Report

Toward a more holistic assessment of NSF performance

This year marks a transition from the assessment approaches used in previous years. While program highlights continued to constitute a large and important part of the evidence we used to make our judgments about *significant achievement* of NSF's strategic goals, the Committee also began to look critically at alternative approaches to assessment and other kinds of evidence. These included additional ways to analyze highlights to draw out substantive themes as well as statistical patterns, which are reflected in this year's subgroup reports. We also prepared case studies of internal evaluation efforts that take a more comprehensive look at the value of specific NSF programs. By exploring these more holistic ways to take a longer and deeper view of NSF's performance, our ultimate goal was to make recommendations that will lead to a richer understanding of the inherent value of NSF's investments in science and the nation.

To begin to explore the feasibility and usefulness of alternative performance assessment frameworks, the Committee this year formed a "Future Assessment Task Group" to look at different ways in which NSF might demonstrate achievement of its goals. The group conducted its review by examining a wider variety of data sources, such as budget, award, and other trends; workshops and reports; program evaluations; evaluation research; career tracking mechanisms; and case studies.

Task Group Members:

Sharon Dawes (Chair)
Senior Fellow, Center for Technology in Government
University at Albany
State University of New York

Diran Apelian
Howmet Professor of Mechanical Engineering and Director, Metal Processing Institute
Worcester Polytechnic Institute

Mary Ellen Sheridan
Associate Vice President for Research (retired)
The University of Chicago

To gain a better understanding of how different approaches might be of benefit, Task Group member Mary Ellen Sheridan conducted a case study of the Information Technology Research (ITR) program to explore how a longitudinal program perspective might enhance performance assessment. Suzi Iacono, Senior Science Advisor, Directorate for Computer Information Science and Engineering (CISE), gave considerable assistance and was present at the AC/GPA meeting to answer questions. ITR was a five-year, $1.116 Billion program to expand the horizons of computing research. It encouraged inter-institutional partnerships across disciplines aimed at high-risk research to design tools for the nation's

cyberinfrastructure. ITR supported more than 500 large and medium-sized grants as well as hundreds of smaller grants. As such, it drew an overwhelming number of proposals and involved nearly every NSF directorate, requiring the invention of a complex internal proposal management and review process that had not existed before. ITR grants produced significant advancements in software design and quality, scalable information infrastructure, high-end computation, and insights into the workforce and socio-economic impacts of technology. It fueled new interdisciplinary areas such as bioinformatics and geoinformatics which are now formal NSF program areas, as well as grid computing and visualization, which continue to impact many other fields.

ITR was not only an investment in computing research, it also provided an opportunity to evaluate the issues associated with collaboration across disciplines and institutions. One evaluation took the form of a research grant to better understand the benefits and problems of such collaborations.[1] The investigators collected social network data from the medium and large grant awards involving nearly 4000 pairs of senior personnel from 475 ITR projects. The research found that successful collaborations had "the right mix" of specialism and diversity and were more likely to occur when investigators had already worked together. The research further found that explicit attention to and resources for coordination, as well as specific coordination activities were associated with success. The most diverse teams were the least productive and those that involved larger numbers of universities were at greatest risk of not publishing; both these findings were explained in large part by insufficient coordination. The investigators concluded that preference for multi-disciplinary research and partnerships bears further examination and that NSF should consider ways to encourage the formative processes of collaborations and the training of scientists to manage them.

Task Group member Diran Apelian organized a group of NSF program officers from the Engineering Directorate who discussed their approach to and lessons learned from assessment. Dr. Allen Soyster, Director of the Division of Engineering Education and Centers (EEC), Dr. Rathindra DasGupta, Program Director for the Industry/University Cooperative Research Centers (I/UCRC) Program, and Dr. James Brown, American Association for the Advancement of Science (AAAS) Fellow, gave presentations that reflected their experiences in assessing the Engineering Research Centers (ERCs) and I/UCRCs, as well as the full range of Industrial

[1] Cummings, J. N., & Kiesler, S. (2007). Coordination costs and project outcomes in multi-university collaborations. *Research Policy, 36(10),* 1620-1634.
Cummings, J. N., & Kiesler, S. (2008). Who collaborates successfully? Prior experience reduces collaboration barriers in distributed interdisciplinary research. *Proceedings of the ACM conference on Computer-Supported Cooperative Work, (November 10-12),* San Diego, CA.

Innovation Partnerships. The presenters emphasized the value of a "systems approach" that focuses on programs and portfolios, rather than individual projects. Such an approach needs to be structured to allow assessment in a more holistic and historical way. A discussion of EEC division-wide assessment identified a taxonomy of concepts and measures that encompass knowledge, human, and economic impacts, as well as the products of their interactions, such as innovation, publications, degrees, and jobs. This taxonomy is illustrated as follows:

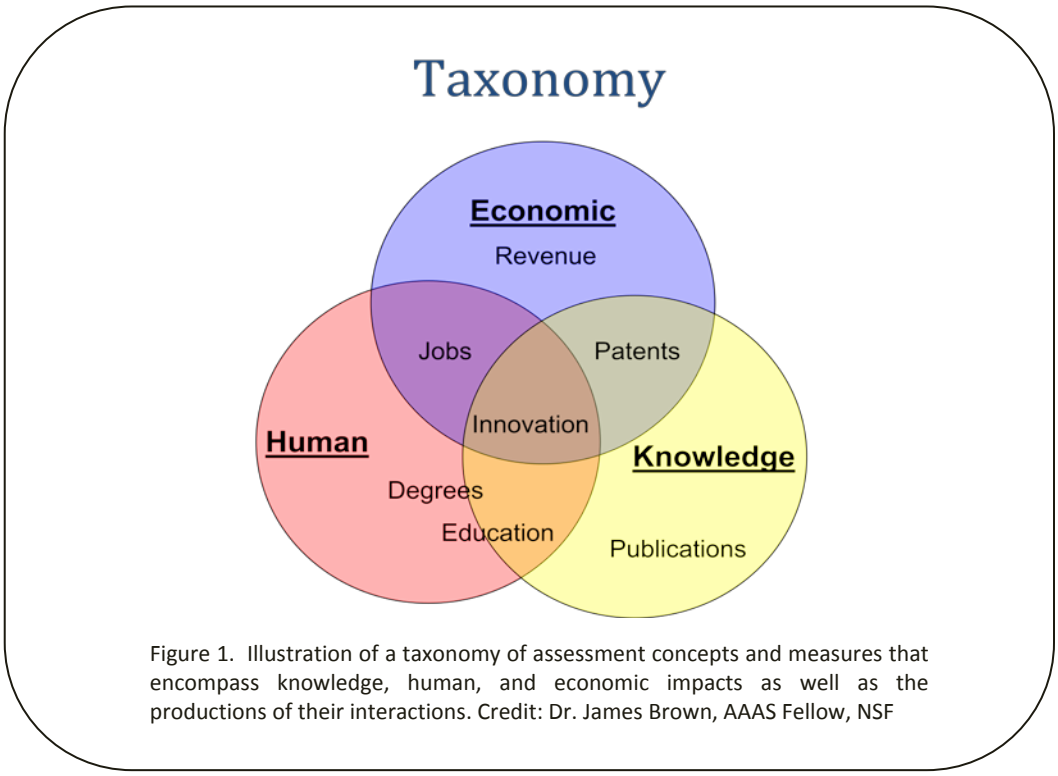

Figure 1. Illustration of a taxonomy of assessment concepts and measures that encompass knowledge, human, and economic impacts as well as the productions of their interactions. Credit: Dr. James Brown, AAAS Fellow, NSF

The Engineering presentation and discussion further highlighted the importance of setting expectations among and building long-term relationships with grantees. The program officers stressed the importance of cultivating Principal Investigators (PIs) as partners in assessment and highlighting the value to PIs of having NSF as a funder and outlet for their work. One example is the publication of an annual compendium of engineering breakthroughs that highlights the work of centers and which provides an incentive to participate actively in reporting and evaluation. These NSF programs encourage grantees to engage in self-assessment and require reporting on scientific, technological, technology transfer, or educational achievements, such as patents and trends in workforce diversity. Throughout the discussion, the speakers emphasized the importance of tracing development over time, linking different measures to the interests of different audiences, and treating assessment as an integral part of the responsibility of PIs and NSF.

The Task Group, along with the subgroup chairs, also tested a different approach to using program highlights this year. The subgroup chairs looked for cross-cutting themes within the highlights assigned to their groups, and the NSF staff provided additional kinds of analysis from the highlights database. As a result, several important themes emerged that had not been visible before. In brief, impressive achievements represented in the highlights seem to signify work that is *integrative* (i.e., it links Discovery, Learning, and Research Infrastructure), *interdisciplinary* (i.e., it draws from multiple fields to generate new fields, questions, or insights); *enabling* (i.e., it lays the groundwork for other discoveries or new directions), and *cross-boundary* (i.e., it builds bridges across the full spectrum of education, across institutions, and between sectors). This different way of considering highlights was aided by the preparation of two "extended highlights" on behavioral economics and polymer science, which demonstrated how NSF investments over time in programs, fields, or subfields have led to significant breakthroughs in our understanding.

In short, NSF investments in science and engineering research and education generate a dynamic set of effects that is much more compelling than has been possible to see through individual highlights alone. This exploration of alternative assessment methods using different kinds of evidence shows how infrastructure investments pay off in discoveries and how the process of discovery embodies learning, both of which enlarge and enhance the STEM workforce. Together with carefully crafted research investment strategies, these investments generate both knowledge impacts and economic impacts of value to the nation and the world. This is the larger story that performance assessment should tell.

Accordingly, the task force offers the following recommendations for constructing a more holistic performance assessment program that will not only better demonstrate the value of NSF investments, but return learning dividends to NSF itself. We understand such a significant shift from current practice represents more than a one-year effort. We therefore look forward to an ongoing discussion in future years about the Foundation's efforts to move in this direction. These recommendations have been endorsed by the full AC/GPA.

RECOMMENDATIONS

1. *Consider an assessment framework that uses multiple measures and methods, applied over various time scales. Use both quantitative and qualitative evidence, including highlights.*
2. *Emphasize the dynamic relationships among strategic goals and outcomes*
3. *Use performance assessment as an opportunity and means to document the strategic value of NSF's science investments to the nation and the public.*
4. *Engage the scientific community as a partner in performance assessment*
5. *Build assessment into the organizational and programmatic infrastructure of NSF.*

1. *Consider an assessment framework that uses multiple measures and methods, applied over various time scales. Use both quantitative and qualitative evidence, including highlights.*

The Committee recognizes that NSF programs represent a portfolio of investments of different size, scope, and duration embodying varying levels of risk. Moreover, these investments represent different kinds of value to different stakeholders. Consequently, we urge the Foundation to consider a matrix of assessment methods and measures that captures a more complete range of impacts and employs simple, easy-to-communicate rubrics and language. Assessment focus areas should include knowledge impacts, people impacts, research strategy impacts, and economic leverage. Alternative methods and measures in each of these focus areas could include:

- Knowledge impacts: tracking results according to themes in the strategic plan and annual budgets, and commissioning internal or external studies on the impact of investments in certain thematic areas of science or on the strategies that NSF uses to encourage cutting edge discovery.
- People impacts: looking at programs and experiences whose purpose is to invest in the STEM workforce, such as Research Experiences for Undergraduates (REUs), Graduate Research Fellowships, the Faculty Early Career Development (CAREER) program, and relationships among PIs and graduate students.
- Research strategy impacts: investigating the efficacy of different kinds of solicitations, proposal requirements, and review processes to address the desire for risk-taking and broader impacts.
- Economic leverage: considering the long-term economic effects of NSF investments in such programs as the Small Business Innovation Research Program (SBIR), ERCs, or via retrospectives on major investment areas such as nanoscience or computer science.

Similarly, the full impact of NSF performance cannot be discerned solely through annual exercises. We believe each of the impact areas listed above can and should be assessed in different ways over multiple time periods. Short term assessments would most likely be annual and largely quantitative, medium term assessments would take place periodically and focus on selected thematic or synergistic topics, and highly selective long-term assessments would take a long retrospective view of major areas of scientific development over time (see Figure 2).

Although many other data sources are salient, we believe highlights will continue to be a useful form of evidence for assessment. We therefore recommend that future highlight databases include the amount and duration of the grants from which the highlights are drawn. We also encourage NSF to continue to prepare idea-centric, and person (career)-centric extended highlights. In addition, more complete contextual information would be useful, including placing NSF's role, contribution, and leverage in the context of the larger science enterprise that includes both private sector and other federal investments in research. Finally, we urge the Foundation to catalog and make use of many existing in-house data sources, such as budget trends and other reports that are useful evidence for performance assessment.

Figure 2. Matrix of assessment areas and timeframes

	Short-term	Medium-term	Long-term
Knowledge			
People			
Research Strategies			
Economic			

2. Emphasize the dynamic relationships among strategic goals and outcomes.

Each strategic goal is important in itself, but far greater value comes from the dynamic interaction among Discovery, Learning, and Research Infrastructure. The Foundation would be well-served to compile and tell that larger story. Committee members suggest thinking in terms of meta-analysis, and multiple, simultaneous outcomes generated by the portfolio, rather than rely so heavily on the accumulated but unconnected results of single investigations.

The Committee also recommends that the Foundation develop and use a taxonomy of assessment concepts and measures that demonstrate the synergy among human, economic, and knowledge impacts. This could be useful not only in the assessment process, but as a way to communicate outcomes with a variety of audiences.

3. *Use performance assessment as an opportunity and means to document the strategic value of NSF's science investments to the nation and the public*

NSF enables scientific advancement as well as education of the STEM workforce and the public. Its investments leverage other resources and encourage risk-taking in the service of discovery and learning. Its work is an important link in a chain of multipliers that deliver the value of scientific investment. The full value of NSF's work may not be measurable in quantitative terms, but it is surely a story worth documenting in the richest possible way.

The Committee recommends that NSF consider selecting a few measures used consistently over a long time period. Suggestions include framing performance in terms of "grand challenges," connecting the past with the future, broadening the concept of scientific infrastructure to include human and knowledge elements, and tracing the global flow of knowledge and scholars through the U.S. research and education enterprise. To communicate effectively, measures and presentation styles need to be customized to address the needs and interests of different stakeholder groups, including an increasingly global and more diverse public.

4. *Engage the scientific community as a partner in performance assessment.*

We endorse the idea that PIs must become partners in performance assessment. We recommend that NSF build assessment requirements into its awards, find ways to widely publicize achievements that return value to PIs and their institutions, and develop ways to track the productivity and creativity of people supported by multiple NSF awards throughout their careers.

In addition, discussions in our Committee meeting returned often to the human dimension. In the context of changing U.S. and world demographics, we urge the Foundation to emphasize the inherent value of broadening participation in science and engineering. To the extent that NSF can foster full participation and career development among women and minorities, it will be advancing not only STEM capability but also the breadth and diversity of scientific inquiry. For example, women's health research has advanced dramatically over the decades since women began to comprise a substantial portion of investigators in that field. A significant increase in currently underrepresented groups within the STEM workforce will not only engage the fastest growing portion of the population in scientific careers, it will open new sets of questions and new approaches to research that are simply not possible when genuine diversity in culture, experience, and world view is not present. Toward that end, it would be very useful to find easier ways to collect diversity data and to prepare and disseminate information about the strategic value of diversity.

5. Build assessment into the organizational and programmatic infrastructure of NSF.

All of the foregoing recommendations rest on the belief that assessment must be treated as a regular ongoing function of the Foundation, with suitable staff and other dedicated resources. As a permanently and professionally-trained staff function, performance assessment can become a process of continual feedback and learning. NSF could begin to capture and share the learning that is already taking place in program- and directorate-level assessments such as those done for ITR and ERCs. Regular, ongoing review of Committee of Visitor (COV) reports and responses to their recommendations should also be part of this more comprehensive assessment function. Such an approach would allow the Foundation to link assessment to changing needs and priorities. It would also allow NSF to consider what it is not doing regarding assessment and communicating impacts, such as finding ways to learn from unsuccessful investments as well as successful ones. Moreover, we urge the Foundation to focus some research energy and resources on assessment by embedding evaluation research into major programs and placing some of the focus of the Science of Science Policy Program on NSF itself. This will be a cultural shift for NSF, but one we believe will bring lasting value to it as an organization.