A Report by a Panel of the

NATIONAL ACADEMY OF PUBLIC ADMINISTRATION
for the National Science Foundation

February 2001

A STUDY OF THE NATIONAL SCIENCE FOUNDATION'S CRITERIA FOR PROJECT SELECTION

Panel Members
James Colvard, *Chair*
Carl Bostrom
Mary Jane England
Sandra Hale

Project Staff

Christopher Wye, *Project Director*
Robert Ross, *Senior Research Associate*

TABLE OF CONTENTS

CHAPTER1: EXECUTIVE SUMMARY

Purpose of Study

On July 17, 1997, the Senate Appropriations Committee report (S. Rept 105-55, FY 98) that accompanied the FY98 VA HUD and Independent Agencies Appropriations Act requested that the National Science Foundation (NSF) contract with the National Academy of Public Administration (the Academy) to review the impact of changes in criteria for the merit review process.[1] This request was reiterated in the FY99 report of the Senate Appropriate Committee (S. Report 105-215) that accompanied the FY99 Appropriations Bill.[2] Through the merit review process, NSF evaluates 30,000 proposals submitted to it annually, out of which it funds approximately 10,000.

The Senate Committee's request grew out of its general concern to ensure accountability and responsibility in funding. To better understand NSF's decision-making process in providing support to scientific research at both the proposal level and the program/priority–setting level, it was necessary to ask such questions as:

- Is it a good process?
- Is the process producing good results?
- Are smaller institutions able to participate?

Methodology

In the design of its new selection criteria, NSF sought to achieve six broad objectives, which are expressed in management guidelines for proposal submission and in published reports:

1. Encourage a broader range of projects to be supported.
2. Seek wider institutional participation (e.g., by smaller as well as larger institutions).
3. Encourage greater diversity of participation in NSF-funded projects by underrepresented minorities.
4. Support projects with positive social impact.
5. Foster the integration of research and education.
6. Simplify the merit review criteria.

These six NSF objectives for modifying or improving the selection process constitute the framework for the analyses used in this study.

The methodology for the study included review of relevant legislation, reports by the General Accounting Office and external review committees; interviews with key personnel in NSF and stakeholders from the scientific and academic communities; comparison of the old and new merit review criteria and selection processes through analysis of a sample of projects funded under both systems and analysis of the behavior and intentions of reviewers in using the new criteria.

Comparison of Old and New Merit Review Criteria

The preexisting selection process rests on four criteria established in 1981. The new process established in 1997 reduces these to two. The following table compares the old and new criteria.[5] *Arrows* indicate the repetition of an element of the 1981 criteria in the 1997 criteria. '*New*' designates elements in the 1997 criteria that have been added. A short analysis of the differences between the 1981 and 1997 criteria may be found in Appendix I at the end of the Executive Summary.

1981 Criteria	**1997 Criteria**
Criterion 1 **Research Performance Competence**	**Criterion 1** **Intrinsic Intellectual Merit**
Capability of proposer →	Qualifications of proposer
Technical soundness of approach →	Well-conceived and organized activity
Adequacy of institutional resources →	Sufficient access to resources
Recent research performance →	Quality of prior work
Criterion 2 **Intrinsic Merit of the Research**	
Leads to new discoveries or advances within own field or impacts other fields →	Advances knowledge and understanding within own field or across different fields
New	Explores creative and original concepts
Criterion 3 **Utility or Relevance of the Research**	**Criterion 2** **Broader or Societal Impact**
Contributes goals extrinsic to research field, basis for new technology →	Disseminates results broadly to enhance scientific and technological understanding
Assists in solution of societal problems →	Proposed activity benefits society
Criterion 4 **Effect on Infrastructure of S & E**	
Contributes to S&E infrastructure: research, education, human resource base →	Enhances infrastructure for research and education: facilities, instrumentation, networks, partnerships
New	Promotes teaching, training, and learning
New	Broadens participation of underrepresented groups (gender, ethnicity, disability, geographic)

Major Conclusions and Recommendations

The following is a summary of the major conclusions of the Academy study of the new NSF merit review criteria.

1. **It is too soon to make valid judgments about the impact and effectiveness of the new merit review criteria. Further study is recommended.**

The new merit review criteria have been in place for too short a period of time to make a valid assessment of their impact on any of the six major objectives NSF has had for instituting them. This is true for both statistical analyses of their impact as well as interpretations of anecdotal perceptions.

The fact that policies and implementation processes within NSF towards achieving the objectives of the new criteria had already begun well before the new criteria were instituted makes determination of a baseline against which to measure the impact of the new criteria difficult, if not impossible. Analysis of project proposal jackets—and Committee of Visitor (CoV) reports, in particular—reflects this. For example, from the evaluation of a sample of proposal jackets from FY97 and FY99, it is not possible to discern any striking difference in the type of project proposals that have received NSF grants after the establishment of the new merit review criteria.

2. **There is a need for quantitative measures and performance indicators to track the objectives of the new merit review criteria.**

Determination of the impact of the new criteria is hindered by the absence of hard data. Therefore, NSF should develop a robust database, adequate quantitative measures, and appropriate performance indicators to determine whether progress toward the objectives for the new merit review criteria is being achieved. Interviews with experts and stakeholders confirm the finding that NSF does not have adequate data to track changes or improvements to encourage a *broader range of institutions* or greater participation by *underrepresented minority researchers*. Even within NSF, a senior statistician in the Office of Integrative Activities (OIA) has concluded that "one cannot at this point assess the impact of Criterion 2 on minorities and women."

It would also be extremely useful for NSF to institute long-term tracking of the effects of its research projects, measuring effects at least 10 years out. The most recent CoV reports strongly reinforce this need for long-term project tracking, and better collection of data relative to the NSF objectives in instituting the new merit review criteria.

NSF has recently proposed a number of new directives to improve the review process (discussed in more detail below). One of these – designing a project reporting format to be consistent with the objectives of the merit review criteria (Option 16) – should be quite useful in tracking what was actually done in a project against what was simply proposed.

3. There is a need to improve the conceptual clarity of the objectives of the new criteria as well as the language used in stating them.

An important premise of rational science is that decisions are made and theories are supported on the basis of empirical evidence. For this reason, asking scientists to speculate on the possible future *broader or societal impact* of a proposal raises a distinct level of discomfort for many reviewers. This discomfort is increased when precise definitions of some of the objectives of the new criteria remain ambiguous.

The conceptual clarity of the new review criteria, therefore, needs to be improved so the criteria better reflect the intentions of NSF for instituting them. This is true of the language of Criterion 2, in particular. Most reviewers interviewed (80%) felt the new merit review criteria had made little or no contribution to achieving NSF's stated objectives in instituting them. While some reviewers (20%) felt these objectives were desirable, many (over 50%) felt the language of Criterion 2 was vague and made the criterion hard to implement. For example, there is ambiguity and a wide range of possible meanings of terms used in Criterion 2 – in particular, "benefits of the proposed activity to society." Interpretations of societal benefit ranged from addressing endemic social or environmental problems to having practical or economic application.

Almost half of reviewers and NSF staff interviewed expressed the view that the objectives of the new two merit review criteria were, in fact, better served by the detail and language of the former four merit review criteria.

A third of reviewers interviewed (33%) were strongly resistant to the objectives of the new criteria – particularly those that sought to address societal needs. Some reviewers felt these goals were not applicable to the kinds of grants they reviewed (largely those in traditional disciplines); other reviewers indicated they simply refused to apply Criterion 2 on the grounds that they did not find considerations of societal impact or infrastructure relevant or meaningful.

4. Virtually all stakeholders interviewed felt that using targeted (set-aside) programs is the best strategy for achieving objectives related to broader impact, particularly the need to improve the participation of underrepresented minorities in scientific research.

Among the objectives of the new criteria related to broader impact, improving the participation of underrepresented minorities is one that is universally valued. This is also an objective that has been given specific emphasis by NSF at least as early as 1992.

There was some division of opinion about whether societal benefit could best be achieved by seeking this as a dimension within all projects or by establishing targeted programs for projects with social relevance. However, most reviewers and experts both within and outside NSF expressed a preference for using targeted (set-aside) programs to improve the participation of underrepresented minorities, rather than forcing these objectives into every project.

5. If NSF wants to make change, it must go beyond simply modifying the language of the review criteria. There is a need to systematically incorporate the objectives of the new criteria into the entire cycle of the review process.

Rewriting the language of the review criteria and restructuring their order is essentially treating only surface-level symptoms and not addressing underlying issues, about which there is considerable diversity of views within the scientific and academic communities. The ultimate differences about issues raised by Criterion 2 are not those of language but of belief. Therefore, establishing a process to ensure genuine attention to the goals of the new criteria throughout the entire review cycle – from proposal submission to proposal review to program management to CoV assessment – is a strategy that will have greater impact than isolated directives focused simply on the language used in announcements and forms. For example, while CoV reports from FY99 discuss the societal impact of proposed research somewhat more frequently than do earlier CoV reports (e.g., those of FY97), they reveal little improved understanding or unanimity about its meaning.

For those reviewers who intend to apply both criteria, the most frequent procedure has been to use Criterion 1 as a cut-off, looking at scientific merit first, and only then apply Criterion 2 to evaluate any remaining proposals. Reviewers who try to apply Criterion 2 as a matter of course in their own evaluation process sometimes find its language unclear. Moreover, even reviewers who try to apply Criterion 2 systematically indicate it plays a more minor role than Criterion 1. Therefore, it is reasonable to infer that Criterion 2 is not being used in a balanced way or with equivalent weight to Criterion 1.

Many experts have also recommended that NSF institute broader-based review panels. This would mean that panels need to be drawn from a wider range of institutions, disciplines, and underrepresented minorities.

NSF Initiatives to Improve the Review Process

NSF has employed several methods to evaluate and improve its merit review process, including administrative reviews and reports from Committees of Visitors. In October 1999, NSF's Office of Budget, Finance, and Award Management developed a number of directives focused on strengthening consideration of Criterion 2.

NSF's intention for the new merit review criteria has been to encourage reviewers to fully address both criteria. However, it has found that to this point there is strong evidence that "many proposers and reviewers are ignoring Criterion 2"[6]. The October 1999 directives were presented as a draft of 16 options to strengthen consideration of Criterion 2. These options were grouped into four categories: (1) proposal development, (2) peer evaluation, (3) development of funding recommendations, and (4) agency management of the merit review process. The options paper was distributed, and in response to comments received, an implementation strategy was drafted in November 1999. The following table lists:

- the first three steps of that implementation strategy
- the four categories under which options were grouped
- the relevant options of the original 16

NSF Directives to Improve the Review Process

Steps	Categories	Options
Step 1 Focus on widely supported options	Proposal Development	**Option 1** Implement new electronic template to ensure integration and diversity language incorporated into all program announcements
		Option 2 Review and revise language in *Grant Proposal Guide*
	Evaluation by Peers	**Option 3** Review descriptive language following each criterion so reviewers understand NSF's intent re "broader impact"
		Option 6 Require reviewers to separately address both criteria by providing separate response sections for each criterion
		Option 11 Discuss importance of both criteria in introduction by Program Officers to panelists
	Development of Funding Recommendations	**Option 12** Review and revise language in *Proposal and Award Manual (PAM)* regarding merit review criteria and process
	Agency Management of the Merit Review Process	**Option 13** Include element in CoV reviews to look at whether both criteria are being addressed
		Option 14 Explicitly address use of both criteria in CoV reporting template
		Option 15 Explicitly address use of both criteria in annual merit review report to NSB
Step 2 Focus attention on options considered less important or requiring consideration to be successfully implemented		**Option 10** Require Program Officer analysis to specifically address both criteria; Division Directors have responsibility for compliance. Develop electronic review form (Form 7) with prompts
		Option 16 Redesign project reporting format consistent with new criteria to track what was expected and what was actually done
Step 3 Assess progress and develop additional options or mechanisms to address areas where insufficient progress made		

These directives focus primarily on Criterion 2 and the need to use both criteria in evaluations. However, merely incorporating language about the integration of research and education, diversity, and societal impact into electronic program announcements may not be sufficient to achieve NSF's objectives. The ultimate differences about issues raised by Criterion 2 are not those of language but of belief, and these need to be addressed directly in appropriate public forums.

The requirement to address both criteria separately in separate response sections is a straightforward, low-tech strategy to encourage separate thinking about Criterion 2. Similarly, as noted earlier, creating project reporting formats consistent with the objectives of the new criteria can be a useful means to track progress towards those objectives. At the same time, since Program Officers have the final say in recommending the funding or non-funding of project proposals, it is surprising that requiring their analysis to specifically address both criteria was not among those options considered as important as others in the first phase. This has been subsequently deemed very important, and NSF has indicated it is being implemented.

Recent CoV reports show that NSF has made improvement in the efficiency goals of the merit review process (percentage of proposals evaluated within six months); however, the effectiveness of the new merit review criteria is less clear. Having two criteria is perceived as simpler than having four. At the same time, responses remain largely free-form, and the use of both review criteria, scientific merit and broad impact, is occurring in less than 50% of the application evaluations.

Recommendations to Expand NSF's Merit Review Process Improvement Initiatives

"If the new Merit Review Criteria are to continue to be used, NSF needs to do a better job educating and coaching reviewers in their use." This quote from the International Programs Committee of Visitors FY99 report reflects a widely held view, and it has several implications for NSF's implementation strategy to improve merit review. The following points present four basic considerations to help NSF increase the likelihood that it can improve the quality of merit review and validate this through performance measurement of the outcomes of that process.

1. **Provide better training for reviewers and Program Officers in the importance of the objectives of the new review criteria for NSF's long-term strategy for improving investments in scientific research.**

The fact that many reviewers either ignore Criterion 2 or in some cases regard it as irrelevant in the review of proposals indicates a need for reviewers to better understand the importance of the objectives of Criterion 2 in NSF's long-term strategy for improving investments in scientific research. New language about integrating research and education, diversity, and social impact must be accompanied by other means to emphasize their importance. This should include training for Program Officers, more explicit guidelines for reviewers, and presentations to major research universities and institutions. The issue of strengthening consideration of Criterion 2 is as much a matter of changing current attitudes as it is simply publicizing the goals of NSF.

2. **Provide better practical instruction for reviewers and Program Officers in how the two new criteria are to be used together.**

Many reviewers perceive Criterion 1 (scientific merit) and Criterion 2 (broader or societal impact) as in competition with each other. Some use Criterion 1 as the base level cut-off, applying Criterion 2 only in cases involving the need to decide among remaining proposals of equivalent scientific merit. Many reviewers (73%) disregard Criterion 2 altogether or simply merge social value into scientific merit. Some reviewers parrot the language of Criterion 2, without making any actual evaluation on the basis of it. Most reviewers feel NSF has not been sufficiently clear about how the two criteria are to be used together.

A number of reviewers indicated that NSF has to give better guidance and instructions to reviewers, including the specific mandate that reviewers address both criteria, assuming that to be an NSF goal. However, examples of where the concerns of Criterion 2 "are and are not relevant" should be included.

3. **Address the intellectual and philosophical issues the objectives of the new criteria raise in appropriate public forums, both to clarify the meaning and application of the objectives, and to generate consensus about their use.**

The concept of *broader social impact* raises philosophical issues for many reviewers – in particular, reviewers who see their task as exclusively one of assessing the intellectual merit of proposals. These issues exist for PIs as well, since many function in roles of both researcher and reviewer. Appropriate public forums in which these underlying issues are debated will eventually accomplish more than attempting to improve understanding solely through one-way directives from NSF.

It is also recommended that NSF encourage Program Officers to take a longer-term view of the goals of scientific research projects and their potential impacts. Program Officers, in making recommendations to award or decline proposals, seek to address NSF's strategic goals. These include "contributions to human resources and institutional infrastructure development, support for 'risky' proposals with potential for significant advances in a field, encouragement of interdisciplinary activities, and achievement of program-level objectives and initiatives." However, Program Officers need a better understanding of the specific processes for the distribution of awards relative to these objectives.

4. **Develop a merit review *process evaluation* strategy based on valid performance improvement principles. This strategy should be supported by both qualitative and statistical data collection methods capable of measuring incremental movement towards achieving NSF's strategic goals.**

Developing an evaluation process to determine the effectiveness of NSF's merit review is an important part of NSF's compliance with the mandate of the Government Performance and Results Act (GPRA). The ability to demonstrate control over this key NSF activity can be a powerful tool to help confirm that Congressional decisions to support its programs were valid.

Just as performance measures answer the question of *what* NSF programs have accomplished, *process evaluation* answers the questions of *why and for what goals* NSF uses merit review and *how* it goes about it. Developing process evaluation for merit review can help Program Officers improve the quality of their performance measures. Robust process evaluation will be supported by statistical data for those outcomes appropriate for measurement by quantification, but will also develop meaningful performance indicators for outcomes requiring qualitative measures to determine their level of achievement and to verify results.

Glossary of NSF Acronyms

National Science Board (NSB)

Office of the Inspector General (OIG)

Office of the Director (OD)
- Office of the Deputy Director (OD)
- Office of Equal Opportunity Programs (OD/OEOP)
- Office of the General Counsel (OD/OGC)
- Office of Integrative Activities (OD/OIA)
- Office of Legislative and Public Affairs (OD/OLPA)
- Office of Polar Programs (OD/OPP)

Directorate for Biological Sciences (BIO)
- Division of Biological Infrastructure (BIO/DBI)
- Division of Environmental Biology (BIO/DEB)
- Division of Integrative Biology and Neuroscience (BIO/IBN)
- Division of Molecular and Cellular Biosciences (BIO/MCB)

Directorate for Computer and Information Science and Engineering (CISE)
- Division of Advance Computational Infrastructure and Research (CISE/ACIR)
- Advanced Networking Infrastructure and Research (CISE/ANIR)
- Division of Computer-Communications Research (CISE/CCR)
- Division of Experimental and Integrative Activities (CISE/EIA)
- Division of Information and Intelligent Systems (CISE/IIS)

Directorate for Education and Human Resources (EHR)
- Division of Educational System Reform (EHR/ESR)
- Division of Elementary, Secondary, and Informal Education (EHR/ESIE)

Office of Experimental Programs to Stimulate Competitive Research (EHR/EPSCoR)

Division of Graduate Education (EHR/DGE)
- Division of Human Resource Development (EHR/HRD)
- Division of Research, Evaluation, and Communication (EHR/REC)
- Division of Undergraduate Education (EHR/DUE)

Directorate for Engineering (ENG)
- Division of Bioengineering and Environmental Systems (ENG/BES)
- Division of Chemical and Transport Systems (ENG/CTS)
- Division of Civil and Mechanical Structures (ENG/CMS)
- Division of Design, Manufacture, and Industrial Innovation (ENG/DMII)
- Division of Electrical and Communications Systems (ENG/ECS)
- Division of Engineering Education and Centers (ENG/EEC)
- Small Business Innovation Research (SBIR)

Directorate for Geosciences (GEO)
- Division of Atmospheric Sciences (GEO/ATM)
- Division of Earth Sciences (GEO/EAR)
- Division of Ocean Sciences (GEO/OCE)

Directorate for Mathematical and Physical Sciences (MPS)
- Division of Astronomical Sciences (MPS/AST)
- Division of Chemistry (MPS/CHE)
- Division of Materials Research (MPS/DMR)
- Division of Mathematical Sciences (MPS/DMS)
- Division of Physics (MPS/PHY)
- Office of Multidisciplinary Activities (MPS/OMA)

Directorate for Social, Behavioral, and Economic Science (SBE)
- Division of Behavioral and Cognitive Sciences (SBE/BCS)
- Division of International Programs (SBE/INT)
- Division of Science Resource Studies (SBE/SRS)
- Division of Social and Economic Sciences (SBE/SES)

Office of Budget, Finance, and Award Management (BFA)
- Budget Division (BFA/BUD)
- Division of Contracts, Policy, and Oversight (BFA/CPO)
- Division of Financial Management (BFA/DFM)
- Division of Grants and Agreements (BFA/DGA)

Office of Information and Resource Management (IRM)
- Division of Administrative Services (IRM/DAS)

Division of Human Resource Management (IRM/HRM)
- Division of Information Systems (IRM/DIS)

Addendum to Comparison of Criteria

The following summarizes the most readily discernible similarities and differences between the old and new merit review criteria.

- All of the elements of the old (1981) criteria reappear at some point in the new (1997) criteria. However, the 1997 criteria add three areas into the evaluation of proposals that are not expressed in the 1981 criteria (designated by 'New' in the graphic):
  - ❖ the creativity and originality of concepts in a proposed activity
  - ❖ the specific intention to promote teaching, training, and learning in addition to advancing discovery and understanding
  - ❖ the objective of broadening participation of underrepresented groups

- NSF has restructured the criteria by essentially placing all of intellectual merit (research performance competence and intrinsic merit of the research) from the 1981 criteria within Criterion 1 of the 1997 criteria. All broader impact and societal objectives have been shifted into the new Criterion 2.

- The immediate effect of this restructuring is to make the broader impact and societal objectives more visible – both to the scientific and engineering communities and to Congress. Many experts interviewed have also interpreted the restructuring of criteria as NSF's intention to make societal goals equivalent to intellectual merit.

- Moving from four criteria to two reinforces NSF's desire to make the criteria conceptually simpler. However, the language of the new (1997) criteria, particularly in Criterion 2, is more abstract and general than that of the 1981 criteria. Therefore, the bifurcation of elements in the 1997 criteria expressed in abstract language creates the possibility of reducing consideration of the broader, societal objectives of the new criteria to a checklist.

- The new (1997) criteria give more emphasis to the importance of promoting networks and partnerships in enhancing the infrastructure for research and education.

The following are additional observations which consider the instructions given to reviewers:

- Reviewers generally had greater freedom in their application of the old (1981) criteria in most areas of proposal evaluation. Instructions to reviewers on use of the 1981 criteria give reviewers particularly great leeway in interpreting and applying their weighting of Criteria 2 and 3.

- For goals beyond the immediate research field of a proposal, the 1981 criteria give greater emphasis to a proposal's utility or practical application; the 1997 criteria appear to give somewhat more emphasis to addressing societal needs.

- Further discussion of the four 1981 criteria in an older document referred to as the Proposal and Award Manual (NSF Manual #10, dated September 15, 1992 – a document NSF indicates is being updated) adds for Criterion 4:

> Included under this criterion are questions relating to scientific and engineering personnel, including participation of women, minorities, and the handicapped; the distribution of resources with respect to institutions and geographical area; stimulation of quality activities in important but underdeveloped fields; and the utilization of interdisciplinary approaches to research in appropriate areas.

However, these additional guidelines do not appear to have been included in the *Information for Reviewers* that accompanied the review forms.

- The instructions to reviewers on use of the new (1997) criteria also give reviewers great freedom of interpretation and application. The FastLane instructions refer to the questions accompanying each criterion as "potential considerations" that might be employed, or "suggestions" not all of which will apply to any given proposal. Further, reviewers are informed that "the criteria need not be weighted equally."

The freedom of application and weighting of the new (1997) criteria may reinforce the unwillingness of many reviewers to apply Criterion 2 at all. In fact, making broader impact and societal objectives more visible may have a reverse effect on those reviewers who reject the idea that evaluations should include consideration of broader, societal goals beyond scientific and intellectual merit.

Footnotes

[1]In 1981, the National Science Board (NSB) adopted four criteria for the selection of research projects: (1) research performance competence, (2) intrinsic merit of the research, (3) utility or relevance of the research, and (4) effect of the research on the infrastructure of science and engineering. In May 1996, the NSB established an NSB-NSF Staff Task Force, charging it to re-examine the merit review criteria and make recommendations on retaining or changing them. On July 10, 1997, NSF announced changes in its merit review criteria (Important Notice No. 121, New Criteria for NSF Proposals). The changes reflected its own analysis and input from the scientific and academic communities.

[2]The charge from Congress for the Academy study of NSF's new merit review criteria was first contained in a Senate report accompanying the National Science Foundation FY 1998 appropriation (July 17, 1997). The Senate Appropriations Committee report (S. Rept 105-55, FY 98) that accompanied the FY98 VA HUD and Independent Agencies Appropriations Act included the following request:

> The Committee is aware that the agency [NSF] recently revised the criteria for merit review of proposals submitted to the agency for funding, and that the criteria now include consideration of the broader applications of the research to be supported. The Committee encourages NSF to examine how the changes in the merit review criteria have affected the types of research the agency supports, and directs the agency to support a review of the new criteria by the National Academy of Public Administration, to be initiated after the new criteria have been in place for 1 year. In addition, the Academy study should address the overall criteria-setting process within the agency, including how the agency identifies areas for new initiatives and measures progress in existing initiatives.

In the FY99 report of the Senate Appropriate Committee (S. Report 105-215) that accompanied the FY99 Appropriations Bill, the Committee said:

> As discussed in last year's report, the Committee expects NSF to contract with the National Academy of Public Administration to review the procedure and criteria for merit review, now that the new criteria have been in place for a year. This study should review the overall merit review process in the agency, as well as examine how the changes in the merit review criteria have affected the different types of research that NSF supports.

[3]The Senate's concern that smaller and lesser-known institutions were not competing well within NSF's merit review system may be a consequence of the perception that, historically, NSF's review process has placed great weight on scientific merit, and less on societal impact. NSF readily acknowledges it has experienced problems in implementing the new criteria, and is trying to address them through management directives and instructions to reviewers. At the same time, with the new merit review criteria having been in place only a little over a year, NSF management also feel it is still too early to do an impact study.

An additional driver for NSF's examination of its merit review system, noted in the FY 1996 Report on the NSF Merit Review System (NSB-97-13), was a fall 1994 Government Accounting Office (GAO) report on peer review at three government agencies (*Peer Review: Reforms Needed to Ensure Fairness in Federal Agency Grant Selection* GAO/PEMD-94-1). This report points out that "although peer review in principle has broad support, there has been a long history of controversy about how it is practiced. The most contentious debates have centered on whether current systems provide fair, impartial reviews of proposals." Among areas of concern, GAO found that junior scholars and women were consistently underrepresented, and that there were problems in the consistency in how review criteria were applied. With respect to the latter, reviewers often "used unwritten decision rules in rating proposals." Partly in response to the GAO report, NSF established a senior-level Peer Review Study Group (PRSG) to examine relevant issues associated with merit review. Subsequently, several task groups of NSF staff examined the efficacy of the process and made recommendations for action. Several stresses and strains on the merit review system were identified. To address these concerns, an external Proposal Review Advisory Team (PRAT) advisory committee was chartered in late FY 1996 to inventory and evaluate current stresses on the system, develop

feasible options for addressing the most important issues, and evaluate options from the perspective of proposers and reviewers. The PRAT met for two days in December 1996, and was to present a report to the Deputy Director by the summer of 1997.

[4]The following are the major elements of the Academy study of the new NSF merit review criteria:

A. The study reviewed relevant documentation including legislative reports and testimony, reports by the General Accounting Office, internal NSF reports on the current and previous merit review systems, reports of external review committees, and reviews from the scientific and academic communities.

B. The study conducted informal discussions with key personnel in NSF to gather background information, decide methodologies for data gathering and analysis, and determine an appropriate sample of projects funded under both old and new criteria. The study then conducted structured interviews with key stakeholders from NSF, Congress, OMB, GAO, experts from the scientific and engineering communities, and academic institutions to seek input on the merit review process, including how the changes in the merit review criteria have affected the different types of research NSF supports.

C. The study compared the old and new merit review criteria and selection processes with respect to their similarities and differences, how NSF identifies new initiatives as a result of changes in the merit review criteria, and how it measures progress.

D. The study assessed the effects of the old and new merit review criteria and selection processes in three ways:

- Through analysis of a sample of projects funded under both systems, identifying trends that can be supported by the data, but also discussing factors limiting the ability to discern valid effects of the new merit review criteria, including the length of time the new criteria have been in effect, the amount of outcome data captured by NSF, and the clarity of measures for some objectives.

- Through analysis of the behavior and intentions of reviewers in using the new merit review criteria, asking reviewers about their judgment process before and after institution of the new criteria, and also examining a keyword study NSF did to estimate how many reviewers used the new Criterion 2.

- Through analysis of the perceptions of stakeholders with special interests in determining whether NSF's new merit review criteria are achieving their intended objectives.

E. The study conducted limited comparisons of NSF's merit review process with that of other similar funding agencies such as NIH, DOD, and the Department of Energy.

[5]Language of the 1981 and 1997 Merit Review Criteria

1981 Criteria

Criterion 1—Research performance competence.
The capability of the investigator(s), the technical soundness of the proposed approach, and the adequacy of the institutional resources available, and the proposer's recent research performance.

Criterion 2—Intrinsic merit of the research.

The likelihood that the research will lead to new discoveries or fundamental advances within its field of science and engineering, or have substantial impact on progress in that field or in other scientific and engineering fields.

Criterion 3—Utility or relevance of the research.
The likelihood that the research can contribute to the achievement of a goal that is extrinsic or in addition to that of the research field itself, and thereby serve as the basis for new or improved technology or assist in the solution of societal problems.

Criterion 4—Effect of the research on the infrastructure of science and engineering.
The potential of the proposed research to contribute to better understanding or improvement of the quality, distribution, or effectiveness of the Nation's scientific and engineering research, education, and human resources base.

1997 Criteria

Criterion 1—*How important is the proposed activity to advancing knowledge and understanding within its own field or across different fields? How well qualified is the proposer (individual or team) to conduct the project? (If appropriate, the reviewer will comment on the quality of prior work.) To what extent does the proposed activity suggest and explore creative and original concepts? How well conceived and organized is the proposed activity? Is there sufficient access to resources?*

Criterion 2—*How well does the activity advance discovery and understanding while promoting teaching, training, and learning? How well does the proposed activity broaden the participation of underrepresented groups (e.g., gender, ethnicity, disability, geographic, etc.)? To what extent will it enhance the infrastructure for research and education, such as facilities, instrumentation, networks, and partnerships? Will the results be disseminated broadly to enhance scientific and technological understanding? What may be the benefits of the proposed activity to society?*

[6]Statement of NSF staff manager.

CHAPTER 2: THE MERIT REVIEW PROCESS

Merit Review Criteria Key Events and Decisions Timeline

1980

1981 National Science Board (NSB) adopts four general review criteria

1984 NSB amends 1977 policy requesting Director of NSF submit an annual report of the NSF proposal review system.

1990

1993 Government Performance and Results Act (GPRA) requiring federal agencies to provide a 5 year Strategic Plan, an annual Performance Plan, and an annual Performance Report.

1994 October: NSB requests reexamination fo 1981 general review criteria

1995 July: Grnat Poliuc manaual (NSF: 95-26 replaces NSF 88-47)

1996 November 20: Discussion Report of Task Force on Merit Review (NSB/MR-96-15)

1997 February 12: FY 96 report on the NSF Merit Review System (NSB-97-13)

1997 March 6: Memo: Analysis of Responses to the NSB/NSF Report on Merit Review Criteria, from Susan Cozzens, Dr., Off. of Policy Support to Paul Herer, Executive Secretary, Task Force

1997 March: Final Report of Task Force on Merit Review (NSB/MR-097-05)

1997 March 28: NSB Approves the use of the new NSF merit review criteria (NSB 97-72) for all proposals reviewed beginning October 1, 1997

1997 July 10: Important Notice Np. 121, New Criteria for NSF Proposals. NSF announces changes in its merit review criteria.

1998 October 1: Grant Proposal Guide (NSF 99-2, replaces NSF 98-2)

1998 Senate Appropriations Committee report (S. Rept. 105-53) which accompanies the FY 98 VA HUD and Independent Appropriations Act directs NSF to engage a NAPA review of the effect of changes in the merit review criteria.

1998 November:L NSF keyword search to estimate percentage of reviews received by NSF that substantively address the second merit review criterion.

1999 Senate Appropriations Committee Report (S. Rept. 105-216) that accompanies the FY 99 Appropriations Bil reiterates the previous year's report that NSF contract with the Academy to review the procedure and criteria for merit review, now that the new criteria have been in place for a year.

1999 March 15: FY 98 Report on the NSF merit Review System (NSB-99-28)

1999 September 10: O/D Staff Memorandum (O/D 99-14) emphasizes importance of Criterion 2 and its connection with GPRA performance goals

1999 September 20: Dear Colleagues letter to PIs and reviewers (NSF 99-172) reiterates O/D 99-14.

1999 September 20: Important Notice to President of Universities and Colleges (Important Notice 125) reiterates O/D 99-14

1999 October 19: Memo from Deputy Director—Guidelines for Advisory Committee Assessment of Directorate Performance for GPRA for FY 1999

1999 October: BFA develops 16 options to strengthen consideration of Criterion 2

2000

2000 February: FY 99 report on the NSF Merit Review System (NSB-00-78 and memo NSB-00-84)

Brief History of the Development of New Merit Review Criteria

The preceding timeline may serve as a point of reference for this brief history of merit review at NSF, with focus on the development of the new merit review criteria. Merit review is a critical component of NSF's decision-making process for funding research and education projects. Almost all of the 30,000 proposals submitted to NSF annually undergo external merit review; however, NSF has resources to fund only about one third. It is through the use of merit review that NSF seeks to maintain standards of excellence and accountability in the funding of scientific research.

In 1981, the National Science Board (NSB) adopted four generic criteria for the selection of research projects:

(1) Research performance competence
(2) Intrinsic merit of the research
(3) Utility or relevance of the research
(4) Effect of the research on the infrastructure of science and engineering

The 1981 criteria addressed only research proposals, because education programs had been eliminated from the budget at that time. However, later in the 1980s, the criteria were adapted to suit education programs as they were reestablished.

As the portfolio of NSF supported projects expanded – to include, in particular, broad education initiatives and research center activities – and as the Government Performance and Results Act (GPRA) emphasized the importance of NSF linking its long-range strategic goals to the results of its investments in science and engineering, the NSB felt an assessment of the appropriateness of the merit review criteria was warranted. In its May 1995 meeting, the NSB stated that reexamining the criteria in light of its new Strategic Plan was a matter of high interest. Following this meeting, the Deputy Directory formed an NSF staff task group on the review criteria. The task group found that the "criteria were unevenly applied by reviewers and NSF staff in the proposal review and selection process." The task group reported that, "The NSB criteria are in need of clarification and should be rewritten." The task group also recommended that options be explored for more effective application of the criteria.

In May of 1996, the NSB established a combined NSB-NSF Staff Task Force on Merit Review, and charged it with examining the Board's generic review criteria, with the purpose of making recommendations to retain or change them, including providing guidance on their use. The Task Force consisted of Dr. Warren Washington (Chair) and Dr. Shirley Malcom, Dr. Eamon Kelly, and Dr. Mary Gaillard from the NSB, and Dr. Mary Clutter, Dr. John Hunt, and Mr. Paul Herer from the NSF staff. The Task Force Discussion Report of November 20, 1996 (NSB/MR-96-15) presented the group's findings, intended not as a final set of recommendations but as a means to stimulate discussion within and outside NSF.

The Task Force met several times for discussion, and reviewed a number of previous studies and surveys, cited in the Report. Some of the more significant findings of the Task Force included the following:

- A cross-section of reviewers in a 1991 NSF/SRI considered the first two of the 1981 NSB criteria (intrinsic merit and PI competence) to be considerably more important than the last two. Less than half said they usually commented on all four criteria.

- Studies by the NSF Office of Policy Support brought to light a number of problems with the 1981 NSB generic criteria, including lack of clarity in wording resulting in idiosyncratic interpretations, the non-uniform application of the criteria (both across the four criteria and across NSF divisions), and the difficulty in applying the criteria to non-research activities such as education and facilities or centers.

Earlier, in February 1996, the NSF Staff Task Group on Review Criteria had recommended the criteria be rewritten, both to make them clearer and to emphasize important attributes such as innovation, clarity of thought, and soundness of approach.

In the Discussion Report, the combined NSB-NSF Task Force recommended two generic criteria to replace the four NSB criteria: (1) *What is the intellectual merit and quality of the proposed activity?*, and (2) *What are the broader impacts of the proposed activity?* Within each criterion were a set of additional questions designed to assist the reviewer in understanding their intent. However, reviewers would address only those elements they considered relevant to the proposal at hand and that they felt qualified to make judgments on.

The Task Force listed a number of advantages of the proposed new criteria:

- "NSF is increasingly asked to connect its investments to societal value, while preserving the ability of the merit review system to select excellence within a portfolio that is rich and diverse. Having two criteria, one for intellectual quality and the other for societal impact, should serve to reveal the situations where proposals have high quality but minimal potential (and vice-versa). Quality will continue to be the threshold criterion, but will come to be seen as not sufficient by itself for making an award."

- "The two new criteria are more clearly related to the goals and strategies in the NSF Strategic Plan. For example, *NSF in a Changing World* states (page 31) that: 'We rely on our proven system of merit review, which weighs each proposal's technical merit, creativity, educational impact, and its potential benefits to society.'"

- "The criteria are simplified by reducing their number from four to two, and are defined for reviewers and proposers by a set of suggested contextual

> elements. Reviewers are asked to describe the proposal's 'strengths and weaknesses' with respect to each criterion using only those contextual elements that they consider relevant to the proposal at hand."

The Task Force also recommended guidance on the use of the criteria. The process issues framing this guidance included (a) the need to maintain flexibility in the application of the criteria because of the great range and diversity of activities supported by NSF, and (b) the need to provide Program Officers with flexibility and discretion in the application and weighting of criteria. An additional process issue was related to the need to modify the NSB generic criteria for projects with special objectives. For example, the CISE Minority Institutions Infrastructure Program Announcement (NSF 96-15) listed nine additional factors that were to be used to evaluate proposals. It was felt that revising the NSB generic criteria would lessen, although not eliminate, the need for special criteria.

The Discussion Report also addressed various options for rating proposals, with the overall goal to encourage reviewers and panelists to provide substantive comments on proposals, not merely "check boxes" on some proposal rating scheme. Whether separate ratings for each of the two criteria or a composite rating were used, instructions and guidance to reviewers would be most important. "The system will be improved only if the reviewer uses the criteria when evaluating the proposal." Thus, the review form and the *Grant Proposal Guide* would need to be redesigned so that both PIs and reviewers understood what was to be evaluated. The Discussion Report provided a sample draft *NSF Proposal Review Form* as well as a synopsis of NSF's strategic plan *NSF in a Changing World* (NSF 95-24) to which outside reviewers should be exposed. Among the long-range goals and core strategies particularly relevant to Criterion 2 were:

- *Promote the discovery, integration, and employment of new knowledge in service to society.*
- *Integrate research and education*
- *Promote partnerships* (including with universities, elementary and secondary schools, and state and local governments)

On October 17, 1996 the NSB approved the release of the Task Force Discussion Report – not as NSB policy but as a proposal for broader discussion inside and outside of the NSF.

On February 12 of 1997 the *FY 1996 Report on the NSF Merit Review System* was released. The report documented the numbers of competitive reviews of proposals (29,953) and awards (8,796), indicating a decline in funding rate to 29% from 34% five years earlier, and a slow increase in proposals from women and minority PIs since 1990. The most frequent method of proposal review was combined mail and panel review (60%, up from 42% in FY 87). The report also noted that NSB and NSF were developing new proposal review criteria, and that the merit review system was undergoing continued examination by internal staff and the external community for ways to improve its efficiency and effectiveness.

The report also noted an additional driver for NSF's examination of its merit review system – a Fall, 1994 Government Accounting Office (GAO) report on peer review at three government agencies. Partly in response to the GAO report, NSF established a senior-level Peer Review Study Group (PRSG) to examine relevant issues associated with merit review. Subsequently, several task groups of NSF staff examined the efficacy and made recommendations for action. Several stresses and strains on the merit review system were identified. To address these concerns, an external Proposal Review Advisory Team (PRAT) advisory committee was chartered in late FY 1996 to inventory and evaluate current stresses on the system, develop feasible options for addressing the most important issues, evaluate options from the perspective of proposers and reviewers. The PRAT met for two days in December, 1996, and was to present a report to the Deputy Director by the Summer of 1997.

In March of 1997, the NSB published its Final Recommendations (NSB/MR-97-05) of the Task Force on Merit Review. The proposed recommendations of the Discussion Report were shared with the science and engineering community through press coverage, contacts among staff, universities, and professional associations, and through a response form on the World Wide Web. NSF received over 300 responses, largely from tenured faculty who had experience with the merit review process.

The Task Force recommended that two new criteria be adopted in place of the four NSB generic criteria (with the sub-questions currently used with each criterion). In addition, the Task Force suggested that the following language be used in a cover sheet attached to the proposal review form, presenting the context for using the criteria:

Important! Please Read Before Beginning Your Review!

> *In evaluating this proposal, you are requested to provide detailed comments for each of the two NSF Merit Review Criteria described below. Following each criterion is a set of suggested questions to consider in assessing how well the proposal meets the criterion. Please respond with substantive comments addressing the proposal's strengths and weaknesses. In addition to the suggested questions, you may consider other relevant questions that address the NSF criteria (but you should make this explicit in your review). Further, you are asked to address only those questions which you consider relevant to the proposal and that you feel qualified to make judgments on.*
>
> *When assigning your summary rating, remember that the two criteria need not be weighted equally. Emphasis should depend upon either (1) additional guidance you have received from NSF or (2) your own judgment of the relative importance of the criteria to the proposed work. Finally, you are requested to write a summary statement that explains the rating that you assigned to the proposal. This statement should address the relative importance of the criteria and the extent to which the proposal actually meets both criteria.*

To implement the new criteria, the Task Force indicated NSF should address issues of the design of proposal review forms (both paper and electronic), training for NSF staff, and revising NSF's proposal preparation guidelines.

The Final Recommendations included a summary of its discussion, on February 19, 1997, of input received relating to the merit review criteria. The following highlights some of the issues raised, along with the Task Force's recommendations:

- A central issue of "weighting or threshold" was raised by approximately 33% of the respondents. The concern was that adopting the new criteria would lead to a decline in NSF's standards of excellence (i.e., "excellent research with OK relevance" would be equated with "OK research with excellent relevance"). Others felt that Criterion 1 was much more important than Criterion 2 and should be weighted accordingly (some even suggesting 90/10). Others criticized Criterion 2 as irrelevant, ambiguous, or poorly worded. Of the options for responding to this issue (including (a) stating the criteria need not be weighted equally, (b) presenting Criterion 1 as the threshold, (c) differentiating criteria for basic research, applied research, and education proposals, and (d) having Criterion 2 address *both* intellectual impact and "broader" impacts), the Task Force recommended the first option (a) because "it does not polarize the research and education communities and can be applied very flexibly."

- For the issue of how to get reviewers to pay attention to the new criteria, the Task Force recommended the cover sheet "PLEASE READ THIS BEFORE BEGINNING YOUR REVIEW!" shown above.

- A substantial number of respondents indicated the question under Criterion 2 dealing with "diversity" was ambiguous. The current language of Criterion 2 reflects the Task Force's recommended rewording.

- For respondents concerned that for much of basic research it was not possible to make a meaningful statement about the potential usefulness of the research, the Task Force recommended rewording the relevant question of Criterion 2 to its current form.

- To eliminate responses of "yes/no" to questions under each criterion, the Task Force recommended the language be changed to make use of such phrases as "To what degree does . . . ?"

The Task Force concluded that the proposed new criteria were flexible enough, in their design and proposed implementation, to be useful and relevant across NSF's many different programs.

On July 10, 1997, NSF announced changes in its merit review criteria in *Important Notice to Presidents of Universities and Colleges and Heads of Other National Science Foundation Grantee Organizations* (Important Notice No. 121). This announcement indicated the NSB had approved new criteria for reviewing proposals, effective October 1, 1997. The criteria and instructions for proposal review were attached to the notice. The

instructions stated the list of potential considerations for each criterion were "suggestions and **not all will apply to any given proposal**."

In 1998, the Senate Appropriations Committee Report (S. Rept 105-53), accompanying the FY98 VA HUD and Independent Agencies Appropriations Act, directed NSF to engage an NAPA review of the effect of changes in the merit review criteria. In November 1998, NSF conducted a brief keyword search (discussed later in this report) to estimate the percentage of reviews received by NSF that substantively addressed Criterion 2. In 1999, the Senate Appropriates Committee Report (S. Rept 105-216) accompanying the FY99 Appropriations Bill reiterated the previous year's report that NSF contract with NAPA to review the procedure and criteria for merit review, now that the new criteria had been in place for a year.

On March 15, 1999 the *FY 98 Report on the Merit Review System* (NSB-99-28) was released. The report documented the numbers of proposals received (28,321) and funded (9280), a decrease of 5.9% from the previous year. Proposals from minority PIs were funded below the NSF average (31% and 33%, respectively). Proposals from female PIs were funded above the NSF average (34% and 33%, respectively). The most frequent method of proposal review continued to be combined mail and panel review (63%). The report noted that in March 1997 the NSB had approved changes to the merit review criteria, becoming operational at the start of FY 98. The report also remarks that Program Officers, in making recommendations to award or decline proposals, seek to address NSF's strategic goals including "contributions to human resources and institutional infrastructure development, support for 'risky' proposals with potential for significant advances in a field, encouragement of interdisciplinary activities, and achievement of program-level objectives and initiatives." There is no data about the distribution of awards relative to any of these objectives.

On September 10, 1999 O/D Staff Memorandum (O/D 99-14) emphasized the importance of Criterion 2 and its connection with GPRA performance goals. The memorandum stated, "We want to ensure that the criterion relating to broader impacts is considered and addressed in proposals and reviews. Program staff have a key role within the community, to stress the importance of both merit review criteria in preparing and evaluating proposals for NSF. The Foundation's GPRA performance plans for FY 1999 and FY 2000 include performance goals for the implementation of the criteria. Our performance will only be successful when proposers and reviewers address the elements of both review criteria appropriate to the proposal, and program officers take the information provided into account in their decisions on awards."

On September 20, 1999 a "Dear Colleagues" letter to PIs and reviewers (NSF 99-172) reiterated the message of O/D 99-14, as did *Important Notice to Presidents of Universities and Colleges and Heads of Other National Science Foundation Grantee Organizations* (Important Notice No. 125), also published on September 20.

In February 2000 the *FY 1999 Report on the NSF Merit Review System* (NSB-00-78) was released. The report indicated it reviewed 28,504 proposals, funding 9112 of them (32%).

The funding rates for proposals from minority PIs were below the NSF average in FY 1999 ("and have been for seven of the past eight years"). The numbers of proposals received from minority PIs has also decreased by 5% since FY 92. On the other hand, the number of proposals received from female PIs has increased by 19% during this same seven year period. The report notes that since 1990 the percentage of proposals reviewed by panel alone has increased from 36% to 47%, mail-only review has decreased from 33% to 18%, and the use of combined mail and panel review has increased from 32% to 35%. These figures possibly reflect a growing number of multidisciplinary proposals and a declining response rate of mail reviewers. The report contains no other discussion relevant to the new merit review criteria except in terms of the Committee of Visitors (CoV) evaluation of GPRA Goal 7: NSF performance in implementing the new merit review criteria. NSF's performance goal for the implementation of the new merit review criteria is stated in the narrative GPRA format. NSF performance is ***successful*** when "reviewers address the elements of both generic review criteria appropriate to the proposal at hand and when program officers take the information provided into account in their decisions on awards," or ***minimally effective*** when "reviews consistently use only a few of the suggested elements of the generic review criteria although others might be applicable." The report characterizes the results as "*largely successful, needs some improvement."* In FY 99, 38 CoV reports rated NSF programs on their use of the new merit review criteria. NSF was rated successful in achieving this goal in 33 CoV reports. In most cases where NSF was not fully successful, it was found that reviewers and applicants were not fully addressing both review criteria.

The report notes that NSF has established guidelines in program announcements requiring applicants and reviewers to address these criteria in proposals and reviews. NSF has recently re-issued guidance to applicants and reviewers, "stressing the importance of using both criteria in the preparation and evaluation of proposals." The following language was added to NSF program announcements and included in the *Grant Proposal Guide*:

> *PIs should address the following elements in their proposal to provide reviewers with the information necessary to respond fully to the above-described NSF merit review criteria. NSF staff will give these elements careful consideration in making funding decisions.*

Finally, the report notes NSF's goals to foster integration of research and education, and to broaden opportunities and enable participation of all citizens – women and men, underrepresented minorities, and persons with disabilities.

Description of NSF Proposal Merit Review and Award Process

The preceding flowchart represents the basic timeline of NSF's proposal merit review and award process. The following narrative briefly describes its major events.

NSF announces various award opportunities through its *Grant Proposal Guide* and through the Program Assistant and the Program Secretary. Research facilities and educational communities become aware of these opportunities. Through individuals who represent these facilities and institutions, proposals are developed that are submitted to NSF in one of two ways: electronically via FastLane on the NSF website, or by ordinary mail. The normal proposal preparation and submission time is approximately 90 days.

Once the proposal is received at NSF, it is distributed to the appropriate NSF Program Officer. There are three basic modes of review (1) mail review, (2) panel review, and (3) combined mail and panel. To preserve confidentiality, proposers are not aware of who reviewers are (although they may suggest reviewers). Reviewers are also not aware of one another except if they serve on a panel. Within each Division, the Program Officer (PO) selects reviewers, whether mail or panel.

In general, 90-95% of proposals receive external peer review. In certain well-identified cases review is waived. These cases include proposals submitted in response to formal solicitations governed by the Federal Acquisition Regulations; proposals to provide goods or services obtained through procurement mechanisms such as contracts and purchase orders; cases which have already been effectively peer reviewed (such as incremental funding amendments, no-cost extensions, certain supplements); cases where peer review is not applicable (such as IGPA awards, safety modifications to ships in the academic fleet, interagency agreements for surveys and data processing); cases where peer review is impracticable (such as international travel grants, awards for logistical support).

There are a number of sources through which NSF obtains ad hoc and panel reviewers. Primarily, these sources include the Program Officer's knowledge of what is being done by whom in the research area, references listed in the proposal itself, and reviewer files in the research divisions. In addition, reviewers may be identified from recent technical programs by the professional societies, recent authors in scientific and engineering journals, computer searches on scientific and engineering abstracts, recommendations from other reviewers, and, as previously mentioned, suggestions by investigators themselves.

The role of review panels is somewhat more comprehensive than that of individual reviewers. Responsibilities of review panels include, in addition to proposal evaluation, concerns for quality control, addressing budget constraints, and balancing research priorities against the need to take risks in new areas of research within the division.

The NSF guidelines for selection of ad hoc and panel reviewers are intended to ensure a selection of experts who can give Program Officers the proper information needed to

make a recommendation in accordance with the NSB criteria for selection of research projects. Reviewers should have special knowledge of the science and engineering subfields involved, as well as a broad knowledge of the infrastructure of the science and engineering enterprise and its educational activities. This understanding relates to societal goals, scientific and engineering personnel, and the distribution of resources to institutions and geographical areas. To the extent possible, reviewers should also reflect a balance of geographies, institutions, and underrepresented minorities.

Historically, certain traditions have tended to arise in scientific disciplines regarding the modality of review used. For example, Physics uses only ad hoc individual mail review, no panel. Panel reviews involve meetings and discussion among reviewers. There is approximately a 50%-60% return rate on ad hoc individual mail reviews. A minimum of three reviews is required. The time period from receipt of proposal at NSF to completion of reviews, analysis, and recommendation is optimally six months.

The Program Officer is responsible for analysis of reviews and recommendations. Thus, the Program Officer is really the final arbiter and the one who recommends whether an award is a given or declined. Reviewers' evaluations, therefore, are not final decision points, although they are generally upheld. However, the review process is ultimately advisory only; the Program Officer makes the final decision. Division Directors (DDs) concur on the Program Officer's recommendations to award/decline.

If the Division Director concurs on an award, the grant is distributed to the proposer's organization via the Division of Grants and Agreements (DGA). If the award is declined, the proposal is returned as inappropriate or withdrawn. The average time of DGA review and processing of an award is 30 days.

Program Officers generally manage roughly 100 awards per year. A typical award runs for three years at an average of $70,000 per year (or $210,000 total). Approximately 30% of submitted proposals are successful and receive awards; approximately 60% of submitted proposals are declined.

Different divisions have different standards for how they approach reviews and grants. There is no forced consistency among divisions. Information on ethnicity/race/gender etc of Principle Investigators is not gathered consistently across divisions. However, special attention is given to conflict of interest issues (reviewers with conflicts of interest with proposers).

Discussion of NSF Strategic Plans

The NSF GPRA *Strategic Plan FY 2000-2005* integrates previous strategic planning activities that resulted in the 1995 *NSF in a Changing World,* the 1997 GPRA Strategic Plan, and the 1998 NSB Strategic Plan. The plan seeks to emphasize outcome goals for its three core strategies of (1) developing intellectual capital, (2) integrating research and education, and (3) promoting partnerships. This section will examine the GPRA plan in light of the several objectives for the new merit review criteria.

The NSF Act of 1950 (PL 810507) defined NSF's mission as *to promote the progress of science; to enhance the National health, prosperity, and welfare; to secure the National defense; and for other purposes.* This authorized NSF to initiate and support basic scientific research, programs to strengthen scientific research potential, science and engineering education programs at all levels, and a science information base for national policy.

Of the six major objectives NSF has for the new merit review criteria,

1. Support a broader range of projects
2. Promote wider institutional participation (e.g., by smaller as well as larger institutions)
3. Encourage of greater diversity of participation by underrepresented minorities
4. Promote projects with a positive benefit to society (societal impact**)**
5. Foster the integration of research and education
6. Simplify the merit review criteria

at least 3, 4, and 5 have some presence in the original NSF mission. This mission sees a need to grow and maintain a scientific workforce, a concern for the societal implications of science and engineering, and an awareness of the importance of educational programs.

At the same time, some of the ambiguity in the language of the new criteria is also reflected in the current NSF strategic plan. Often, this ambiguity is a consequence of NSF's desire to pursue multiple directions simultaneously. For example, the plan states "We support a portfolio of investments . . . promoting disciplinary strength while embracing interdisciplinary activities."

Language relevant to the first objective (*broader range of projects*) appears more frequently in the current plan: "Our investments promote the emergence of new disciplines, fields, and technologies." NSF supports academic institutions that are "crucibles for expanding the frontiers of science and engineering knowledge, and educating the next generation of scientists and engineers."

The FY 2000-2005 Strategic Plan defines its strategy for pursuit of its mission in terms of three outcome goals:

1. People: *to develop a diverse, internationally and globally-engaged workforce of scientists, engineers and well-prepared systems.*
2. Ideas: *to support discovery across the frontier of science and engineering, connected to learning, innovation and service to society.*
3. Tools: *to provide broadly accessible, state-of-the-art and shared research and education tools.*

The language of these outcome goals is generally consistent with the language of the new merit review criteria.

Outcome Goal 1 (People) addresses review criteria objectives 1, 3, and 5. Specifically, NSF intends to "use all aspects of NSF activity to enhance diversity in the science and engineering workforce, with particular attention to the development of people who are beginning careers in science and engineering." NSF also plans to help "increase the Nation's capacity to educate teachers and faculty in SMET [scientific, mathematics, engineering, and technology] areas" and "foster innovative research on learning, teaching, and organizational effectiveness." An underlying goal seems to be to try to make science and engineering fields attractive to underrepresented minorities by using educational programs to establish meaningful career paths.

Outcome Goal 2 (Ideas) addresses review criteria objectives 1 and 4. Specifically, NSF seeks to "take informed risks" and "provide long-term support for new and emerging opportunities within and across all fields of science and engineering." NSF also strives to "foster connections between discoveries and their use in the service to society."

Outcome Goal 3 (Tools) is primarily oriented with review criteria objective 5 and the infrastructure which supports the integration of research and education.

In discussing its strategy to guide the priorities expressed in the three outcome goals, the Strategic Plan indicates that "NSF's merit review process is the keystone for award selection. All proposals for research and education are evaluated using two criteria: the intellectual merit of the proposed activity and the broader impacts of the activity on society. Specifically addressed in these criteria are the creativity and originality of the idea, the development of human resources, and the potential impact on the research and education infrastructure." NSF defines three *core strategies* as the mechanism to achieve its outcome goals.

The first, ***develop intellectual capital*****,** seeks "investments that tap into the potential evident in previously underutilized groups of the Nation's human resource pool" (review criteria objective 3). The second, ***integrate research and education***, includes investing in "reward systems that support teaching, mentoring and outreach" (review criteria objective 5). The third, ***promote partnerships***, is a general strategy for collaboration, both

between disciplines and institutions, and among academia, industry and government. This strategy potentially impacts review criteria objective 2 (*wider institutional participation)*.

A theme that emerges from the Strategic Plan is the linkage of merit review criteria objectives. For example, the goal of establishing a competent twenty-first century scientific workforce is connected to the need to improve SMET education from pre-kindergarten through elementary and secondary to undergraduate, graduate, and continuing professional education levels. This strongly connects review criteria objectives 3 and 5, since many of the educational programs are targeted toward underrepresented minorities. NSF seeks "a more inclusive and globally engaged SMET enterprise that fully reflects the strength of America's diverse population." NSF feels that "at present, several groups, including underrepresented minorities, women, certain types of institutions, and some geographic areas, perceive barriers to their full participation in the science and engineering enterprise. NSF is committed to leading the way to an enterprise that fully captures the strength of America's diversity."

There is a similar, though somewhat less discussed, linkage between review criteria objectives 1 and 4, where support of more risky, innovative, or interdisciplinary projects is defended in the context of the goal of promoting projects with a positive benefit to society.

The Strategic Plan cites a number of "critical success factors" to manage its activities towards NSF's goals. The first and perhaps most important (Factor 1) is **operating a credible, efficient merit review system.** NSF states that "the merit review system is at the very heart of NSF's selection of the projects through which its outcome goals are achieved." Among the implementation strategies to achieve this are to:

- Regularly assess performance of all aspects of the merit review system, comparing its efficiency, effectiveness, customer satisfaction and integrity against similar processes run by other organizations.
- Promote the use of both merit review criteria (i.e., *intellectual merit* and *broader impacts*) in the evaluation of proposals.
- Develop alternative mechanisms for obtaining and reviewing proposals and evaluating their potential for use in determining NSF's investments.
- Reduce the burden on proposers and reviewers while maintaining the quality of decision processes, by increasing award size and duration.

It is not clear to what extent NSF has in fact conducted a "regular assessment" of its merit review system by "comparing its efficiency, effectiveness, customer satisfaction and integrity against similar processes run by other organizations." It is also not clear that NSF has conducted systematic assessments of its own review processes other than through the Committee of Visitor reports, ongoing general NSF staff review, and its high level annual *Report to the National Science Board on the National Science Foundation's Merit Review System*. (Issues of NSF and comparative assessments of the review process are discussed in more detail in 7.2). Finally, it is not clear to what extent NSF has *successfully* promoted the use of *both* merit review criteria. The data from this study

strongly suggests that reviewers either do not use Criterion 2 (*broader impacts*) at all, or, if they do, do not use it in a *balanced* or *equivalent* manner to Criterion 1 (*intellectual merit*).

The Strategic Plan's discussion of NSF's performance assessment process warrants some additional comment. NSF begins discussion of this process in Appendix 2 of the Strategic Plan with a disclaimer. "The challenge of performance assessment for NSF is that both the substance and the time of outcomes from research and education activities are largely unpredictable." While this is true of the substance of scientific research, it is not true of NSF's own processes, including the merit review process. NSF indicates that "OMB authorized NSF to use alternative format performance goals for our outcomes in research and education. This approach allows for human judgment to consider both quantitative and qualitative information on performance and to weigh that information in a balanced assessment. NSF uses the descriptive performance goals in our management process through a combination of internal self-assessment and review by independent external panels of experts and peers." The Plan goes on to state that "For the three outcome goals, NSF's performance will be considered successful when, *in the aggregate, research or education results reported in the period demonstrate that significant and sufficient progress has been made toward realizing the long-term outcomes and implementing the planned strategies."*

Assessment of goal achievement is performed by external groups of peers and experts at several stages in the grant award cycle. This consists of:

- Applicant and Grantee Information/Merit Review. This is the standard proposal merit review process discussed earlier in this chapter.
- Program Evaluation by Committees of Visitors (CoVs). External experts review each program every three years and report on the integrity and efficiency of the processes for proposal review and the quality of results of programs. Representative CoV reports are discussed in Chapter 5 this analysis.
- Directorate Assessment by Advisory Committees. Directorate advisory committees review internal self-assessments, CoV reports, available external evaluations, and annual directorate performance reports, judging program effectiveness, and describing strengths and weaknesses. The advisory committees' reports are reviewed by NSF management, which integrates recommendations into the NSF Annual Performance Report.

The Strategic Plan indicates NSF has "several mechanisms in place for producing valid and reliable performance measures and assessments." This includes a data quality improvement program on the NSF corporate database, and strategic planning discussions by advisory committees every three years.

Discussion of Instructions to Reviewers

Old Merit Review Criteria

The instructions to reviewers on use of the old four review criteria first state the criteria, then offer the following as guidance.

> Criteria 1, 2, and 3 constitute an integral set that should be applied .....in a balanced way to all research proposals in accordance with the objectives and content of each proposal. Criterion 1, research performance competence, is essential to the evaluation of the quality of every research proposal; all three aspects should be addressed. The relative weight given Criteria 2 and 3 depends on the nature of the proposed research; Criterion 2 intrinsic merit, is emphasized in the evaluation of basic research proposals, while Criterion 3, utility or relevance, is emphasized in the evaluation of applied research proposals. Criterion 4, effect on the infrastructure of science and engineering, permits the evaluation of research proposals in terms of their potential for improving the scientific and engineering enterprise and its educational activities in ways other than those encompassed by the first 3 criteria.

Observations

- These instructions appear to give reviewers enormous freedom of interpretation and application, particularly in the weight of application of Criteria 2 and 3, which *depend on the nature of the proposed research.* The instructions also leave it up to the reviewer to determine the *consequences* of the nature of the proposed research in the evaluation.

- The instructions here do not raise some of the specific concerns of the new review criteria, for example: *broadening the participation of underrepresented groups.*

- The sense in which the instructions encourage consideration of the *broader impacts* of the proposed research is largely in terms of the practical *utility* of the research. While Criterion 3 itself does speak of *assisting in the solution of societal problems,* the instructions for Criterion 3 emphasize its application *in the evaluation of* ***applied*** *research.*

- Further discussion of these four criteria in an older document provided by OIA in ascii text and referred to as the *Proposal and Award Manual* (NSF Manual #10) and dated September 15, 1992 (a document in the process of being updated), adds that

  > Criterion 3 also relates to major goal oriented activities that the Foundation carries out, such as those directed at improving the knowledge base underlying science and technology policy, furthering international cooperation in science and engineering, and addressing areas of national need.

It also adds that for Criterion 4

> Included under this criterion are questions relating to scientific and engineering personnel, including participation of women, minorities, and the handicapped; the distribution of resources with respect to institutions and geographical area; stimulation of quality activities in important but underdeveloped fields; and the utilization of interdisciplinary approaches to research in appropriate areas.

These instructions do counter some of the above observations that concerns present in the new review criteria were not present in the old criteria. However, these additional guidelines did not appear to be included in the *Information for Reviewers* that accompanied the review forms.

New Merit Review Criteria

The instructions to reviewers on use of the new two review criteria (in FastLane) provide following guidance before stating the criteria:

> Please provide detailed comments on the quality of this proposal with respect to *each* of the two NSF Merit Review Criteria below, noting specifically the proposal's strengths and weaknesses. As guidance, a list of potential considerations that you might employ in your evaluation follow each criterion. These are suggestions and not all will apply to any given proposal. Please comment on only those that are relevant to this proposal and for which you feel qualified to make a judgement.

After stating the criteria and the *considerations* for each, the instructions add:

> Please provide an overall rating and *summary statement* which includes comments on the relative importance of the two criteria in assigning your rating. *Please note* that the criteria need not be weighted equally.

Observations

- In slightly different ways, these instructions also give reviewers enormous freedom of interpretation and application, particularly in the weight of application of each criterion. Reviewers are free to determine which *are* and which *are not* relevant to any given proposal. In addition, reviewers are evidently free to weight the criteria on a basis of their own determination, since *need not be weighted equally* is accompanied by no other specification.

- The explanation of how to understand the meaning of each criterion is deliberately opened to subjective interpretation (or even non-application) since the explanations of each criterion are characterized as only *potential considerations* or *suggestions.*

- In saying that not all considerations will apply to any given proposal, and that the criteria *need not be weighted equally*, the instructions essentially give reviewers the license to not apply Criterion 2 at all.

Additional Documents that Include Instructions to Proposers and/or Reviewers

- The *User-Friendly Handbook for Project Evaluation* (NSF 93-152) states it is intended to help PIs and project evaluators think practically about evaluation. However, it primarily geared to internal evaluation of projects, and does not contain anything specific to the new review criteria. OIA also indicates that NSF 93-152 is outdated, published in FY 1993, long before the new criteria were established. They will look into updating it.

- The *Grant Proposal Guide* (NSF 99-2), published October 1998 and replacing NSF 98-2 contains a brief discussion of the new review criteria. It indicates that proposals are carefully reviewed usually by three to ten persons outside NSF who are experts in the particular field represented by the proposal. Before listing the criteria, the *Guide* also says that the criteria are designed to be useful and relevant across NSF's many different programs, however, NSF will employ special criteria as required to highlight the specific objectives of certain programs and activities. No further specification or description of these objectives or programs is given.

- The *Guide* goes on to say that following each criterion are potential considerations that the reviewer may employ in the evaluation. These are suggestions and not all will apply to any given proposal. Each reviewer will be asked to address only those that are relevant to the proposal and for which he/she is qualified to make judgments. This last set of remarks appears to leave it quite indeterminate to the proposer which considerations relating each criterion will be applied as well as in what way they will be applied.

- An updated *Grant Proposal Guide* (NSF 00-2) repeats the above language, but adds discussion in two specific areas, indicating that PIs should address these elements in their proposal to provide reviewers with the information necessary to respond fully to the merit review criteria. It also states that NSF staff will give these elements careful consideration in making funding decisions. The first area concerns the integration of research and education. The updated *Guide* states that one of the principle strategies in support of NSF's goals is to foster integration of research and education through the programs it supports at academic and research institutions. These institutions provide abundant opportunities where individuals may concurrently assume responsibilities as researchers, educators, and students, and where all can engage in joint efforts that infuse education with the excitement of discovery and enrich research through the diversity of learning perspectives. These remarks fairly clearly convey NSF's perception of the value of careers in science that embody *both* research and learning. The second area concerns integrating diversity into NSF programs, projects, and activities. Here, the updated *Guide* states that broadening opportunities and enabling the participation of all citizens – women and men, underrepresented minorities, and persons with disabilities – are essential to the health and vitality of science and engineering. NSF is committed to this principle of diversity and deems it central to the programs, projects, and activities it considers and supports. Again, this guidance to the proposer more clearly conveys the importance of projects that permit

diversity among those who participate. Another NSF publication, the *User-Friendly Handbook for Mixed Method Evaluations* (NSF 97-153) published in August, 1997, also strongly reflects NSF's attention to diversity in participation.

- The Directorate for Education and Human Resources, Division of Undergraduate Education produced *A Guide for Proposal Writing* (NSF 98-91). This *Guide* is specifically oriented towards proposals in research and education. Since reviewers are drawn from two- and four-year colleges and universities, secondary schools, industry, foundations, and professional societies and associations, the *Guide* urges proposal writers to learn the general demographics of the reviewers for the program for which they are submitting proposals. The majority of proposals submitted to the Division of Undergraduate Education are evaluated by panel review. The *Guide* goes on to elaborate on each of the two review criteria in the context of proposals oriented toward undergraduate education by listing questions typically raised in the review process.

  Questions relating to Criterion 1, intellectual merit, include:

  - ❖ Does the project address a major challenge facing SMET undergraduate education?
  - ❖ Does the project have potential for improving student learning of important principles of science, mathematics, engineering, or technology?

  Questions relating to Criterion 2, broader impacts, include:

  - ❖ Are the results of the project likely to be useful at similar institutions?
  - ❖ Does the project effectively address . . . objectives [such as to] increase the participation of women, underrepresented minorities, and persons with disabilities; provide a foundation for scientific . . . literacy?

CHAPTER 3: EVALUATION OF SAMPLE PROJECT JACKETS FY97 AND FY99

Summary of Findings

From a statistical standpoint, with a sample of 50 project jackets (25 from FY97 and 25 from FY99), it is not possible to draw any strict quantitative conclusions. Rather, this section identifies themes that emerge with respect to each of six major objectives in NSF's instituting the new merit review criteria.

For obvious reasons, the FY97 project jackets will not contain direct evaluations of the new merit review criteria. However, since there are many indications that work towards the general objectives of the new criteria began long before the new criteria were instituted – at least to the extent that NSF was concerned about these matters – it makes sense to put questions about these objectives to the FY97 jackets.

- Overall, it is not possible to discern any striking difference in the type of proposals that received NSF grants after the establishment of the new merit review criteria.

- Most grants appear to be awarded to PIs (typically white males from well-established universities) who have received previous awards for related research.

- Generally, there is little effort to provide any explanation of the social impact of the proposed research – its importance in a broader framework – both before and after FY97. Occasionally, in the FY99 grants, a reviewer will take seriously the instruction to evaluate the proposals using *both* Criteria 1 and 2. Even then, however, the explanations of how the research will impact society or provide greater access to underrepresented minorities and women seem forced. Most reviewers prior to FY99 either ignore Criterion 2, dismiss it as irrelevant, or find that the research, to which they nonetheless give high ratings, does little to address the goals expressed in Criterion 2.

- In FY97, reviewers almost as frequently as in FY99 addressed social impact, contribution to education, and minority opportunities. However, on the basis of this sample of project jackets, it appears to be true – regardless of whether reviewers address the goals of Criterion 2 or not – that these goals do not strongly influence the awarding of grants.

- The meanings of *broader range of project* and *positive societal impact* are particularly difficult to apply in the case of many areas of primary scientific research. It can simply be the case that *anything* we learn about our environment – whether about grasshoppers, salamanders, DNA, or fungi – is worthwhile and may lead to (often unexpected) social impact or benefit. The benefit may come long after the research when a project's results are examined in a wider context.

- It can be argued that questions about *positive societal impact*, *broader range of projects and institutions*, and the like must be put to Congress as well. What does Congress regard as relevant to a *positive societal impact*? What are specific areas of society about which it is concerned? Is Congress concerned about the proportion of public moneys spent on projects with some real potential for social impact or which contribute in some meaningful way to our knowledge of ourselves and our environment? Congress may have felt that it was important for NSF, PIs, and reviewers to begin to think in some meaningful way about how these proposals and funded areas of research contribute to the improvement of life, and to recognize that because public money is funding these projects, they need to demonstrate some kind of accountability in a manner accessible or understandable by the public. Interviews with the Senate Appropriations Committee staff have been unrevealing about Congress' underlying interests or motives.

- Similarly, it would be worthwhile to identify to what extent Congress is interested in making research grants available to different kinds of institutions, and to new or minority researchers. While Congress does not speak with one voice, enacted legislation has encouraged NSF in this direction.

- No data adequately tracks "broader range of institutions" or "underrepresented minority researchers." A senior NSF statistician in OIA concludes that one cannot assess the impact of Criterion 2 on minorities and women, although the numbers have generally been going up. His personal view is that the small economic payoff of getting a Ph.D. may be a factor in NSF not been getting many proposals from minorities. In 1980, 350 out of 21,208 or 2% of total proposals submitted were from minorities; in 1990, 1169 out of 28,840 or 4%; in 1999, 1422 out of 28,502 or 5 %. The proposal load rose from 21,208 in 1980 to 28,840 in 1990 and since then has remained relatively static. Submissions by females were 1307 or 6% in 1980; 4004 or 14% in 1990; 5296 or 19% in 1999. A difficulty is that information about race or ethnicity is captured only by certain divisions, and even then it is only for PIs, not the other participants in the research project.

- On the basis of the limited sample of the study, it does not appear to be true that grants are being awarded to a broader range of institutions or to minority researchers or to researchers without a track record of having received (usually numerous) grants. The indicia evaluated by reviewers in assessment of the PI (e.g., publications, receipt of previous grants, infrastructure of the institution) do not promote and may even preclude selection of "new" investigators or different types of institutions.

The following are the principle questions addressed in evaluating the project jackets.

1. Does this project represent a **broader range of project** supported?

2. Does this project represent a **wider institutional participation**? (e.g., by a smaller as well as larger institution)

3. Does this project indicate encouragement of a **greater diversity of participation by underrepresented minorities**?

4. Does this project represent have a specific **positive societal impact**?

5. Does this project foster the **integration of research and education**?

6. For FY 99 projects, in what ways do reviewers attempt to use Criterion 2?

Findings on Specific Questions

1. Does this project represent a **broader range of project** supported?

In the 25 jackets from **FY97**, this topic was discussed in 13. In 10 cases, the project proposed was a renewal of a project, a continuation of previously funded or similar research, or a request for equipment already being used in related projects. It was often difficult to distinguish between a proposed project that truly represented a new or broader range of project and a proposed project that the reviewer simply felt was important or in need of study. In some cases, *broader* was associated not with content area of research but with the methods used in research. In some case, *broader* was discussed in terms of a concern with other PIs doing similar work. In other cases, *broader* was associated with a reviewer's perception of the uniqueness of some aspect of the project. Since a *broader range of project* is meaningful in terms of a particular area of scientific inquiry, reviewers need specific criteria and guidelines with which to identify it. Such criteria might include such descriptors as:

- for the first time
- an area not yet studied
- challenges an existing theory
- collecting new evidence in support of
- extends the application of a technology to
- addresses questions not answered in previous research

In the 25 jackets from **FY99**, this topic was discussed in 10. In 4 cases, the project proposed was a continuation of previously funded or similar research. In 7 cases reviewers expressed concerns about similarity between the proposed project and other existing research projects. In asking more rigorously whether a proposed project is different from other identified research, reviewers may be giving greater attention to this objective, albeit in the context of a negative judgment. Concerns expressed included whether a proposed project was sufficiently creative or original as compared to work

going on under another grant, whether some new equipment or methodology could actually improve measurement capabilities, whether the same questions were being answered in similar studies. In several cases, however, reviewers pointed to the fact that a proposed project would investigate a specific area of need *not* being addressed elsewhere (e.g., "so few studies on scientists and engineers with disabilities"). In a few cases, a project proposing the application of a particular new technology (e.g., "new high-resolution infrared spectroscopic techniques") to investigate certain phenomena would elicit a reviewer evaluation of "cutting edge."

2. Does this project represent a **wider institutional participation**?

In the 25 jackets from **FY97**, this topic was discussed in 8. This objective is particularly difficult to assess because the criteria for *wider institutional participation* are essentially undefined. Where relevant to the nature of the proposed project (e.g., development of educational resources), reviewers note possible benefit to smaller regional colleges or colleges which have historically served underrepresented minorities. *Wider institutional participation* is also at times interpreted to refer to collaborative efforts with other labs or research groups, or interdisciplinary activities (e.g., among physicists, microbiologists, geologists in the use of certain equipment or systems, or among teachers or educators and primary researchers). An important although oblique interpretation of *wider participation* occurs when a project has indirect participatory or education benefits to students or the general public. For example, an astronomical research observatory may serve many audiences – professional, semi-professional, students, and the interested public. Or the results of a particular research project may have useful dissemination to secondary education institutions through state or national programs. A final interpretation of *wider participation* is noted when reviewers feel a certain project may lead to additional experiments involving research facilities at other institutions, where the project itself involves multiple institutional participation, or where there will be collaboration between private industry and institutional research facilities. All of the categories of *wider participation* need to be identified and statistically tracked.

In the 25 jackets from **FY99**, this topic was discussed in 6. Observations by reviewers about *wider institutional participation* for proposals in FY99 do not differ substantially from those in FY97. To the extent that there are projects specifically geared towards educational benefits, a number of reviewers note that proposals will provide these benefits (e.g., that elementary school students will visit the college for hands-on science activities). One project emerges from a two-year junior college; another involves participation by students with disabilities in an AAAS program. However, it is not possible from this limited sample to identify any trends towards an overall wider institutional participation.

3. Does this project indicate encouragement of a **greater diversity of participation by underrepresented minorities**?

In the 25 jackets from **FY97**, this topic was discussed in 8. Forms identifying race/gender/ethnicity of PIs were included in slightly over 50% of the sample.

Particularly where relevant to the nature of the project (e.g., education of the next generation of environmental scientists), proposals indicate the participation by, even the "recruitment" of, minority students; some projects specifically target ethnic groups (Hispanics, Native Americans). Some of the reviewer discussions about *diversity* of participation speak to the particular qualifications or track records of the PIs (adding *intellectual diversity* to research in the field) rather than their specific ethnic or racial background. However, where a PI has the opportunity to engage underrepresented minorities in the project it is often noted. Discussion of "talented female investigators" within a field appear to receive somewhat more attention – possibly because there are significantly more females (as a class of *underrepresented*) than ethnic or racial minorities, possibly because there may be female reviewers who give particular attention to other female investigators. No direct correlation has been made to the gender/race/ethnicity of reviewers. A few reviewers speak to the impact of a PI on training undergraduate and graduate scientists from a diverse population. Projects geared to specific ethnic or racial groups (e.g., improving the performance of African American children in mathematics) typically have PIs representative of that group. Proposals from institutions serving underrepresented groups (e.g., as Morgan State represents a historically black college/university or Cuyamaca Community College serves "at risk" students) typically involve a project in scientific learning or pedagogy as much as research independent of its educational dimension. POWRE (Professional Opportunities for Women in Research and Education) proposals are clearly directed at addressing issues of representation within the scientific workforce.

In the 25 jackets from **FY99**, this topic was discussed in 11. In FY99 there generally appears to be more attention to involvement of underrepresented minorities. Forms identifying race/gender/ethnicity of PIs were included in approximately 70% of the sample. Reviewers appear more deliberate in pointing out the ability of female or minority PIs to build a career and reach out to other underrepresented minorities in science, serving as encouragement for members of those groups to consider careers in science. The opportunity for projects to include outreach to elementary and secondary students, especially in low-income areas, also appears more prominent in the project descriptions themselves. Projects focused on educational programs for ethnic and racial minorities are clearly in a recruitment mode, even serving as a "formalized mechanism of recruiting minority graduate students." At the same time, a sizable percentage of reviews of project proposals contain no discussion of underrepresented groups whatsoever. It remains an unanswered question whether attention to underrepresented minorities should constitute a dimension (at least to some degree) in every proposal or should be restricted to project proposals specifically targeted to address the encouragement of underrepresented groups to enter scientific research. Proposals from universities that have made a deliberate attempt to increase the numbers of minority graduate students (e.g., as the Department of Biology at UCSD) typically make this known in the proposal; some proposals indicate they will track participation by underrepresented minorities in classes or labs given by the PI. As a matter of coincidence, one proposal from the FY99 sample itself directly addresses the issue of majority-minority representation in electoral districting, and the hypothesis that oddly shaped districts to support minority representation may depress political involvement and participation (the PIs turn out to be

white males). In some areas of scientific research – particularly electrical engineering and computer science – data on foreign participants may be useful. The same would be true for students with disabilities, if that data is not already collected.

4. Does this project have a specific **positive societal impact**?

In the 25 jackets from **FY97**, this topic was discussed in one way or another in all 25. First, it is virtually impossible to distinguish between reviewer comment which extols the scientific impact of a project from that which specifically praises its societal impact. From the standpoint of a reviewer or PI operating from within the context of a particular scientific discipline, *scientific merit* and *social value* simply merge. Among the reasons a person chooses to work in a particular scientific discipline is the perception that this area of scientific inquiry has societal value. Exceptions to the natural merging of scientific and social value occurs – in a certain sense – with projects specifically targeted at improving some area of society. For example, a project relating to the education of the next generation of environmental scientists "provides the opportunity for scientists and engineers to be aware of the social, legal or economic implications of their work . . . leading to the development of solutions to global environmental problems." In these cases, *societal impact* is directly built in to the project.

More typically, social impact may be an indirect or long-term benefit of basic research. For example, a project examining the properties of grasshopper communities may in the long run also be "invaluable in programs aimed at controlling damaging densities of grasshoppers." Clearly, the criteria for *positive societal impact* need to be identified in a precise way. Projects specifically targeted at social impact should probably constitute a separate category. Neither of these approaches will likely resolve the natural tendency to merge perceptions of *scientific merit* and *social value*. However, establishing clearer criteria for what constitutes *societal impact* will help establish this goal of Criterion 2. Is this goal any or all of the following:

- to fund *more* projects with direct social impact
- to build a *dimension* of societal concern into all projects
- to raise *awareness* of the importance of social impact in the thinking and planning of PIs when framing a project

But even basic research (e.g., a study of electronic excitations in low-dimensional systems of optically detected resonance), insofar as it is likely to bring increased fundamental insight may "yield significant new physics in several major areas." Thus, if a project is perceived as leading to an expanded knowledge base within a broad discipline (e.g., physics) or extending to other disciplines (e.g., engineering), it will naturally be perceived as having a *societal impact* within the scientific community. Some reviewers tend to use terms which are abstract and global (e.g., "will lead to better science"), reducing the possibility of understanding what the specific impact of a project may be. In some cases, the nature of the proposed project stretches the boundaries of *societal impact* to the broadest theoretical or philosophical level (e.g., a project on binary star formation and the evolution of stellar clusters and planet formation). In other cases, *societal impact*

may very much be interpreted to mean *practical application* (e.g., as in the acquisition of a scanning force microscope which may have practical applications to such industries as shipbuilding, mining, or agriculture). In still other cases, *societal impact* may be interpreted to mean *economic benefit* (e.g., as in a computational project on massively parallel processors which "could have a multi-billion dollar impact on a number of US oil companies"). In yet other cases, *societal impact* may be localized to a particular science (e.g., physiology) but one of particular *human* interest (e.g., improving our medical knowledge of DNA metabolism, or understanding genetic inheritance). Finally, there are cases where *societal impact* has proximate meaning for some natural community but broader implications for the interaction between human and natural communities (e.g., investigating root growth at the ecosystem level).

Reviewers do point out where there are weaknesses in proposals, for example, which give "little indication of how work in other areas will be influenced." In targeted proposals, the *societal impact* is typically obvious (e.g., where "presumably better performance by African American children in math will ultimately lead to more access to mathematics-related professions" or where "the objective is to encourage students to 'tinker' and learn"). All of this discussion reinforces the need to provide clear and distinct criteria for *societal impact.*

In the 25 jackets from **FY99**, this topic was also discussed, in one way or another, in all 25. The observations pertaining to this topic for FY97 generally apply equally to proposals from FY99 with a number of interesting differences. (1) For FY99 proposals *societal impact* is understood in terms of a somewhat longer future perspective (e.g., "the importance of study of cluster-assembled magnetic nanostructures will only increase in coming decades"). It is possible this may be a result of greater awareness of GPRA objectives, which encourage organizations to set 5 and 10 year improvement goals; however, there is no specific evidence to support this interpretation. (2) In a similar vein, reviewers sometimes point to *societal impact* in terms of a global perspective (e.g., "produce scientists who will lead the development of solutions to global environmental problems"). (3) There is a somewhat greater emphasis on interagency cooperation (e.g., "builds on work supported by NASA and DOE"), sometimes for cost-reduction as well as specific societal objectives. (4) There is also somewhat greater emphasis on *societal impact* interpreted as supporting *new* areas of disciplines or *breakthroughs* in technologies (e.g., "mobility in protein NMR structures is an emerging area in biophysics"). (5) Although it does not necessarily distinguish them from FY97 proposals, some FY99 projects are geared to address very specific environmental concerns (e.g., "particular concern with the recent toxic bloom of Pfiesteria on Maryland's Lower Eastern Shore and the diverse perspectives that stakeholders have on its causes and consequences"). (5) Finally, there is an interesting interpretation of *societal impact* in a few reviewer comments that point to the ability of a project to "influence the planning of future large-scale systematic projects." Attention to the systems nature of research is consistent with the emphasis on interdisciplinary and interagency cooperation.

It must also be acknowledged, however, that reviewers' observations of "potential for high impact" are often very abstract and general, conveying more a sense of providing

broad and inclusive lip-service to Criterion 2, rather than making specific connections between the proposal and the objectives of the criterion. This is balanced by a slightly higher incidence of reviewer comments that criticize a proposal for its lack of attention to *societal impact*. However, there was little evidence that reviewers used Criterion 1 and 2 in a *comparative* way (vs. *independently*) in their assessments. As in the case of targeted proposals from FY97, a number of FY99 proposals also specifically target research into the social dimensions of science (e.g., the extent to which engineering as a profession has been motivated by patriotism vs. lifestyle considerations; a project examining alliance-building heretical social movement organizations; a project likely to "contribute to realization of what possibilities exist for bright students with disabilities").

5. Does this project foster the **integration of research and education**?

In the 25 jackets from **FY97**, this topic was discussed in 16. The most frequent interpretation of this objective was expressed by reviewer observations that a proposal involved the participation of graduate students, and, in some cases, "enhanced an ongoing teaching program" or internship program. The opportunity for involving graduate students was regarded as important for attracting new individuals into a particular field. A second level of integration typically involves postdocs working with senior scientists at some specialized research facility. Postdocs especially represent the "lifeblood of scientific research" – a key to the development of human resources. One rather loose interpretation of *integration of research and education* was noted in terms of plans for reports of results of research "to be disseminated to the field in a timely manner" – in some cases involving the internet. Another interpretation of *integration* pointed to the benefits of collaborate and cross-disciplinary studies.

As in the case of targeted proposals addressing other objectives of Criterion 2, some proposals (e.g., developing new curricula in mathematics) were directly intended to "lead to the professional development of science teachers in schools", particularly those serving underrepresented minorities, community colleges, or at-risk youth. Others projects (e.g., establishment of a virtual reality laboratory for engineering education) were concerned with pedagogy as much as the content of the discipline. Some combined cutting edge technology and pedagogy in ways to address specific existing weaknesses in the teaching of science (e.g., a proposal on electronic homework and intelligent tutoring on the web in the context of course delivery tools for large enrollment science classes).

In the 25 jackets from **FY99**, this topic was discussed in 14. In general, reviewer observations about projects' *integration of research and education* are consistent with those of FY97. As in the case of topic 4 above, there is a somewhat greater incidence of reviewers who are critical of proposals for *not* adequately addressing issues of integration of research and education. A reviewer notes, for example, that a proposal "does not discuss university infrastructure or whether younger members of the faculty and students will receive training on the new equipment." Similarly, concern for "communicating knowledge between scientists and non-scientists" receives more attention. By and large, the primary interpretation of *integration* remains the involvement of graduate students, the ability to attract graduate students to an area of research, and maintaining ongoing

teaching or lab programs. Where projects involve labs or specific research facilities, *integration* is sometimes extended to mean using the facility as a "training center" – most often where training is in the use of equipment or software. Interpreting *integration* in terms of publications also occurs, particularly where joint authorship of papers by senior researchers and students is customary. Reviewers also will especially note PIs who do a "good job in taking the results out into the community." To a lesser extent than *societal impact*, but still necessary, this objective should distinguish among several categories of *integration of research and education*. In particular, the goals of enhancing existing educational programs and developing the scientific workforce should be distinguished from the broader goal of improved communication between the scientific and non-scientific communities.

6. For **FY 99** projects, in what ways do reviewers attempt to use Criterion 2?

Among the 25 proposals in this sample, the overwhelming number of reviewers did not use Criterion 2 at all. A rough assessment of the sample was made in terms of three categories:

(1) Does the reviewer attempt to use Criterion 2 as intended?
(2) Does the reviewer not use Criterion 2 as intended or parrot the language without evaluation?
(3) Does the reviewer not use Criterion 2 at all?

Approximately 16% of reviewers attempted to use Criterion 2 as intended. Approximately 11% of reviewers largely parroted the language of Criterion 2 but did not make any actual evaluation on the basis of it. Approximately 73% of reviewers did not use Criterion 2 at all.

Clearly, reviewers would benefit from specific instructions directing them to use **both** Criterion 1 and 2, and to discuss *in their review* how they went about using the criteria.

Discussion of NSF Keyword Search: *Estimation of the Percentage of Reviews Received by NSF that Substantially Address the 2nd Merit Review Criterion*

This brief study from NSF's Office of Integrative Activities (OIA), dated May 18, 1999, began in November, 1998, using a search scheme to estimate the number of proposal reviews received by NSF that substantially addressed the new Criterion 2. Based on a small sample of reviews, the study identified eight terms that had a high incidence of use in reviews addressing the Criterion 2. The Budget Division then devised a query that searched for seven of these terms in all reviews received through FastLane from 1/1/98 through 9/30/98. This amounted to approximately 17,000 reviews. OIA then adjusted the search results for the rate of use of each of the terms based on the small sample. OIA's conclusion was:

> **The search results indicate that 48% of proposal reviews substantially address the 2nd merit review criterion.**

Some questions about the validity of this conclusion are discussed below. First, however, it will be useful to summarize, step-by-step, the methodology used, as presented in the report on the study.

OIA first selected 16 terms for the initial search of a small sample of reviews. The terms appear in the descriptive text of the new Criterion 2 as issued by the Director in *Important Notice No. 121: New Criteria for NSF Proposals*, dated July 10, 1997. These 16 search terms were:

Impact	Infrastructure
Discovery	Research and Education
Understanding	Facilities
Teaching	Instrumentation
Training	Network
Learning	Partnership
Participation	Benefit
Underrepresented	Society

The small sample of 1,123 reviews represented five programs in four directorates: BIO, GEO, MPS, and SBE. OIA selected these five programs because they had a history of receiving reviews via FastLane. For each review in the sample, it was determined which of the 16 search terms were used. The reviews were read to identify those that substantively addressed Criterion 2. For each set of reviews containing a search term, OIA then calculated the number of reviews that addressed Criterion 2 as a percentage of the total number of reviews in which the term was used. Eight of the terms ranged from 76% to 100%. The other eight terms ranged from 0% to 59%. OIA then used the top eight terms for a broader search:

Impact	Infrastructure
Training	Research and Education
Participation	Partnership
Underrepresented	Society

The Budget Division assisted by devising a database query that searched for these terms in a larger sample of reviews. The sample consisted of 16,661 reviews received through FastLane from 1/1/98 through 9/30/98. The query identified 7,845 reviews that contained one or more of the seven terms ("participation" was inadvertently omitted). 82% of reviews containing one or more of the search terms substantively address Criterion 2. In addition, the reviews in the sample that address Criterion 2 and contain one or more of the seven search terms represent only 80% of the total number of reviews in the sample that address Criterion 2.

The study estimated that 8,041 reviews, or 48% of the total reviews in the large sample, substantively addressed Criterion 2. This was based on the following calculation:

> 82% (percentage of reviews containing one or more of the search terms that address Criterion 2) of 7,845 reviews (number of reviews in the large sample that contain one or more of the search terms) = 6,433 reviews (number of reviews that contain one or more of the search terms and address Criterion 2).
>
> 6,433 reviews = 80% of 8,041 reviews (total number of reviews in the large sample that address Criterion 2). 8,041 reviews = 48% of 16,661 reviews (total number of reviews in the large sample).

Discussion

It is not the primary purpose of the Academy study to either challenge or defend the conclusion drawn by this NSF keyword search. The basic reason is that it is not clear what meaning or implications the conclusion – as it stands – may have. It is appropriate that NSF would draw at least the initial inferences from this data.

At the same time, some general questions can be raised about the conclusion:

- Does the occurrence of these keywords validly indicate that Criterion 2 is being addressed by reviewers? Reviewer interviews conducted as part of the Academy study have suggested that many reviewers gave lip service to the language of Criterion 2 without substantially applying it.

- A simple keyword search cannot discern the meaning or intention behind the use of certain keywords, only their occurrence. An additional process to make inferences from their occurrence (and an underlying theoretical argument to support this process) would therefore be needed to draw any more substantive conclusions than *these keywords occurred in 48% of proposal reviews.*

- The criteria for identifying *reviews that substantively address the 2nd criterion* are not stated.

- Some keywords are likely to occur or be used in the context of applying Criterion 1. This might be the case for *impact*, *infrastructure* (particularly if developing or expanding infrastructure were directly a part of the proposal), *partnership, participation,* and possibly others.

- Many keywords have multiple denotations (reference), some of which might apply to Criterion 1 or to something other than the objectives of Criterion 2. For example, *society* might refer to the "society" of species (birds, insects) being investigated rather than to human society. *Impact* could refer to almost anything.

- The best evidence for the intentional application of Criterion 2 would be the reviewers' own statements about how and in what ways they applied it. This information, put in correlation with the keyword search, would make the study far more robust.

CHAPTER 4: EVALUATION OF COMMITTEE OF VISITORS REPORTS

Summary of Findings

From a statistical standpoint, with a sample of 26 CoV reports (13 from FY97 and 13 from FY99), it is not possible to draw any strict quantitative conclusions. Rather, this section will identify themes that emerge with respect to each of six major objectives in NSF's instituting the new merit review criteria. For obvious reasons, the FY97 CoV reports will not contain evaluations of the new merit review criteria. However, as with the evaluation of project jackets, since work towards the objectives of the new criteria began long before the new criteria were instituted, it makes sense to put questions about these objectives to the FY97 CoV reports. One should expect that implications for the objectives of the new merit review criteria would typically appear as weaknesses in the old criteria and review process. There is a much greater discussion of the review process in CoV reports. Further, since CoV reports assess performance within divisions and discipline areas, there is a somewhat wider range of evaluation displayed.

- Overall, there is no striking difference in the conclusions that may be drawn from analysis of the Committee of Visitor reports and the project jackets discussed earlier. However, there is a stronger call for hard data to provide evidence of the degree to which NSF is achieving its goals in instituting the new merit review criteria. This is particularly true of data concerning geographic, gender, ethnic, institutional, and other types of desired diversity in project participation.

- Grants continue to be awarded to PIs who are typically white males from well-established universities. Among under-represented minorities, women have made the greatest progress in improving percentages of participation in funded projects.

- The structure provided for evaluating programs by the GPRA questions in the FY99 reports appears to have resulted in a spotty increased attention to statistics, and the use of a somewhat more formal evaluation process. However, judgments continue to be made through example rather than hard data to substantiate them. For example, a CoV report might typically claim a program had been "successful" in meeting GPRA Goal 3 but provide no data to support that claim.

- In areas relevant to several objectives for the new merit review criteria, vagueness in terminology resulted in a lack of shared understanding and interpretation about what achieving those objectives would mean. For example, *broader range of projects* was in need of several sub-categories to avoid its being reduced to simply "cross-disciplinary" projects. *Positive societal impact* suffered from many, sometimes inconsistent, interpretations in meaning. The criteria for *wider institutional participation* remain largely undefined.

- Some discussions of objectives that were clearer with respect to their meaning and value at times became obfuscated because of differing strategies about *how to achieve* those objectives. This was especially true of the process to encourage *greater diversity of participation by underrepresented minorities*. The lack of statistical data, other than in targeted programs, generally perpetuated differences about strategy and implementation.

- Generally, the CoV reports reveal little improved understanding or agreement about the meaning of the *societal impact* of proposed research. While FY99 CoV reports discuss *positive societal impact* more frequently than those of FY97, this does not entail that it is any more clearly understood. The fact that the majority of both PIs and reviewers still do not even address the new Criterion 2 is some indication that its use, at this point, remains minimally effective. Some reviewers interviewed flatly indicated they had no intention of using Criterion 2 at all, even as a second-level judgment among projects of equivalent scientific merit.

- The relatively short period of use of the two new merit review criteria is the basis for the greatest difficulty in making definitive assessments of their impact. This view is reflected in many CoV reports.

- The CoV reports – particularly those from FY99 – contain many valuable suggestions and recommendations about the merit review process. These should be captured and tabulated as important "voice of the customer" input.

The following are the principle questions addressed in evaluating the CoV Reports:

1. Does the process result in **broader range of projects** supported?

2. Does the process result in **wider institutional participation**? (e.g., by a smaller as well as larger institution)

3. Does the process encourage **greater diversity of participation by underrepresented minorities**?

4. Does the process result in projects with a **positive societal impact**?

5. Does the process foster the **integration of research and education**?

6. Is the process **simpler?**

Comments on Specific Questions

1. Does the process result in **broader range of projects** supported?

In the 13 CoV reports from **FY97**, the topic of *broader range of projects* was discussed primarily in terms of proactiveness in supporting "innovative high quality research" or

the success rates of new faculty submitting proposals. Consistent with this sense of *broader*, committees noted divisions (e.g., Integrative Biology and Neurosciences) that did not allow "diverse, innovative, or multidisciplinary proposals to fall through the cracks" and that were particularly "adept at anticipating the needs of emerging areas, and in taking reasonable risks in supporting innovative and exploratory research." On the other hand, particularly where a program was intended to encourage new initiatives, committees noted instances (e.g., the Instructional Materials Development Program) where there was "relatively little emphasis on high-risk proposals" and in which the CoV "encourages the IMD to increase the number of shorter-term, prototype projects that are more high-risk." Beyond occasions for either praise or criticism, CoVs tended to acknowledge an "appropriate level of high-risk proposals" (e.g., Cell Biology), with the recommendation that "outcomes be carefully tracked."

Using a somewhat more formal procedure (structured by the GPRA questions) for evaluating programs, the 13 CoV reports from **FY99** typically discussed *broader range of projects* in terms of "flexibility of the award process which allows support of a number of different mechanisms to foster basic science, multidisciplinary approaches, and new initiatives" (Neuroscience). CoVs made judgments made through examples rather than hard data about the strength of a program or cluster (e.g., Neuroscience) to meet GPRA Goal #1 (discoveries at and across the frontiers of science and engineering). At the same time, some increased attention to statistical data was also apparent ("Forty-five percent of new awards were awarded to new investigators . . . which is greater than the NSF-wide goals of 30%"). In some cases CoVs recommended that targeted programs "be established for first time investigators" (with the assumption that this might lead to a broader range of supported projects). In some cases (e.g., Materials Research), CoVs found that the GPRA goal that 30% of funded proposals must go to PIs who had never before received funding from NSF was "impractical within the realities of many disciplines." There appeared to be a loose connection between disciplines that were currently active and vital research areas and higher percentages of funding of higher-risk projects and first time PIs.

As in other areas, *broader range of projects* was in need of distinct sub-categories to clarify its possible meanings and intentions. Many reviewers and CoVs simply equate cross-disciplinary projects as *broader* or *innovative* by definition. However, other reviewers deliberately resist the notion that interdisciplinary projects are necessarily innovative, and approach them with considerable skepticism.

2. Does the process result in **wider institutional participation**? (e.g., by a smaller as well as larger institution)

This topic received virtually no discussion in the sample of CoV reports from either **FY97** or **FY99**, except in the context of the general distribution of proposals funded "with regard to geography, race, gender . . . size of institution." Thus, *wider institutional participation* was generally interpreted in terms of the concerns of topic 3 (whose discussion follows). The observations applied directly to institutions are also applied to the process for selecting reviewers. For example, "Geographic distribution of panelists

reflects the population distribution of neuroscientists in the country." For the most part, CoV reports express very general, sweeping evaluations, indicating that the precise criteria for *wider institutional participation* remain largely undefined. We never get inside the meaning of "type of institution." For example: "selection of reviewers based on geography, type of institution and group representation are adequate" (Bioengineering and Environmental Systems); "no evidence of any imbalance . . . by any characteristic such as geography, type of institutions" (Upper Atmospheric Research Section). Comments that are critical of current practices get one step closer to what the objective is looking for: "less than satisfied with the balance of small institutions" (Information and Intelligent Systems Division). But clearly, the goals of wider institutional distribution of funded projects and reviewer participation need still further definition.

3. Does the process encourage **greater diversity of participation by underrepresented minorities**?

In the 13 CoV reports from **FY97**, the topic of *diversity of participation by underrepresented minorities* was discussed with the frequent call for "articulating measurable objectives for the number of minority PIs" (Integrative Biology and Neurosciences). The meaning of *underrepresented minorities* is relatively clear, even where the best strategy for addressing their needs is not. Difficulty in measuring this objective is strongly tied to the absence of statistical data: "one obstacle in quantifying the effectiveness of this program is the lack of data on the number of women and minorities that submit proposals to the regular proposal stream" (Oceanography/Applied Ocean Science). Existing demographics of participation by women and minorities within a given field is a limiting factor, and contributes to the difficulty in evaluating program success. "The CoV is unable to determine whether the small number of applicants reflects the demographics in the field or that the program does not meet the needs of minorities" (Oceanography/Applied Ocean Science). Most divisions are concerned to encourage more minorities to participate in science and engineering, but there is less agreement in how to achieve this. Frequent recommendations include establishing targeted programs for underrepresented minorities and better publicizing of existing programs. Targeted programs such as POWRE (Professional Opportunities for Women in Research and Education) get mixed reviews – several CoVs observing that its "review process is not consistent."

Proactive efforts to engage underrepresented minorities are acknowledged: "the current PD . . . attended the 1996 meeting of the National Society of Black Engineers to encourage their involvement in NSF" (Chemical and Transport System Division); however, the effectiveness of such initiatives is unclear. There is a general sense that programs have greater success attracting proposals from women than from racial or ethnic minorities: programs "support a growing number of female, under-represented minority, and young investigators, although the number of minority PIs is still seriously limited" (Integrative Biology and Neurosciences). Limitations in reaching minorities often lead to the recommendation for targeted programs: e.g, "funded Centers and Projects . . . that address the needs of underrepresented groups" (Advanced Technological Education Program). However, there is not sufficient tracking of the results of such

nascent efforts to encourage the participation, hiring, and mentoring of people from under-represented groups (cf., Instructional Materials Development Program).

Evaluations of CoVs indicate that it is somewhat easier to ensure gender, racial, ethnic, and geographic diversity among *reviewers*, to the extent that such diversity exists within a discipline or research area; however, this typically entails "expanding the pool of reviewers." Even where they do not meet them, some divisions (e.g. Physics) have goals for support of women and minority PIs; some divisions do not appear to have specific goals in this area, although the concern is generally one of high priority. Again, even where CoV reports indicate a program "succeeds in assembling full representation of the diversity of American science among its panelists, ad hoc reviewers, and grantees" or exibits "no significant imbalances in the distribution of awards" (Cell Biology Program), it is not precisely clear what this *success* means.

Statistical data of some sort is more frequently used in targeted programs. For example, the CoV report for the Centers of Research Excellent in Science and Technology (CREST) program, which focuses on minority institutions and their ability to tap an important part of the human resources pools, notes that "racial minority groups currently constitute about 20% of the general US adult population, but only 11% of doctorate program recipients." The NSF report *Science Indicators* tracks the change in enrollment of students with different ethnicity for the years 1980 to the present. Here, "the increase in Black or Hispanic students in graduate studies is far below the 64.9% increase in non-resident aliens that occurred during this period." Hence, the CREST program is targeted at 8 mostly Black and Hispanic universities. The CoV also urges that it "should be made clear in the solution and to proposers and referees that the goal of the program is to produce more minority students earning doctorates." Similarly, another program – Graduate Research Traineeships (GRT) – emphasizes that "strong partnerships between historically black colleges and universities and majority institutions should be encouraged."

In the 13 CoV reports from **FY99**, the topic of *diversity of participation by underrepresented minorities* was a matter of frequent concern. The discussion in FY99 CoV reports is largely consistent with that of FY97 reports, with some areas of greater emphasis. First, there is a greater awareness of existing levels of diversity within a given field: "Geographic distribution of panelists reflects the population distribution of neuroscientists in the country. Gender distribution of panelists is approaching parity. There is a need to increase the numbers of underrepresented minorities serving as panelists; this is unlikely to occur until a larger pool of minority scientists becomes available." (Neuroscience) Second, there appears to be greater awareness of NSF initiatives in this area: NSF is seen as promoting an environment that encourages underrepresented minorities to participate in all stages of the review process. As a result, minority scientists and students are encouraged to consider careers in research and to communicate with NSF program officers and NSF-supported researchers. Third, the ability to meet diversity goals for women is the area of greatest success. Fourth, there is a somewhat more frequent use of statistical or tracking data: "48% of the proposals were awarded to women, 9% were awarded to underrepresented ethnic groups, and 1% were

awarded to individuals with disabilities" (Neuroscience). This may be a consequence of the directive of GPRA outcome Goal 3 which seeks to foster a diverse, globally oriented workforce of scientists and engineers resulting from NSF investments. Nevertheless, CoV reports still emphasize "it is difficult to quantify outcomes" even where they believe the goal is being achieved successfully. Hence, the call for more rigorous tracking: "NSF needs to develop more advanced information systems to find reviewers for proposals. It also needs to collect data concerning the geographic location, gender, ethnicity, etc. for reviewers . . . If NSF wishes to ensure geographic, gender, ethnic and institutional type of diversity in its reviewer pool, then it needs to put in place processes to monitor and assure that this exists" (Anthropological and Geographical Sciences).

Many CoV reports indicate that programs "find it frustrating to try to increase the participation of members of underrepresented groups." One suggestion is that "consideration be given to searching out individuals from the underrepresented groups who hold PhDs and serve as faculty who have not received NSF funding as PI. Efforts should be made to involve these individuals in the review process to help ensure diversity among reviewers . . . " (Anthropological and Geographical Sciences). Another recommendation is that "program announcements need to include specific language to reflect the foundation's concern for participation by underrepresented groups and its support for new investigators."

A number of CoV reports begin to address the specific impact of the new merit review criteria on minorities, but point out that "we do not have much information about underrepresented groups." In fact, there is "little evidence on which to base *any* statement regarding participation of underrepresented groups." Thus, where CoVs claim a program has been successful in meeting GPRA Goal 3 – "Successful . . . CGS cluster has been active in promoting diversity in awards" (Civil and Mechanical Systems of Engineering), there is no data to support the claim. In another example, "the CoV was not provided with field-specific data on gender and minority distributions, so it was not possible to make other than a qualitative statement" (International Programs). In cases of panel review, where "the diversity . . . is excellent", it would seem relatively easy to capture diversity data. There is also a need to capture diversity data beyond the participation of PIs: "award recipients should be encouraged to report gender, ethnicity, and citizenship of investigators and graduate students. Methods should be identified to capture and model these data in validated ways to extend conclusions beyond the incomplete datasets" (Bioengineering and Environmental Systems).

Finally, even where a division (e.g., Astronomical Sciences) captures statistical data

> "Statistics from an American Astronomical Society survey in 1990 show that 11.2% of the membership was female, 92.5% white, 3.7% Asian/Pacific Islander, 1.1% Hispanic, 0.3% African American, and 0.1% Native American Indian. In comparison, from figures provided from final reports of REU sites funded between 1988 and 1994, the 273 students were 44% female, 4% Hispanic, 3% African American, and 7% Asian/Pacific Islander."

while these numbers are higher than the general population of working astronomers, it is still "difficult to judge trends in female and underrepresented minority population . . ."

Again, success is often dependent on targeted programs. "The Division Director created a Reserve fund for redirecting allocations. One clear outcome of this initiative is that the number of female PIs has increased by 50% in just 3 years. The number of underrepresented minority PIs is still discouragingly small." (Materials Research) However, targeted programs can evidently go both ways: "Between 1998 and 1999, the merit review process changed, in part because of the decision to phase out the Graduate Minority Fellowship Program" (Graduate Fellowship Program). Even in targeted programs, there remains an issue of NSF providing adequate information to minority communities: "concern expressed . . . in the low number of applicants from Hispanic Serving Institutions (HSIs), Tribal Colleges and Historically Black Colleges and Universities (HBCUs). The NSF is encouraged to intensify its efforts in increasing applications from these institutions" (Graduate Fellowship Program). It also must be asked against what standard is the "use of the new merit review criteria successful"? Only on the basis of such a standard can one interpret the data that the "number of minority Fellows supported in the GRF mechanism rose from 41 in 1998 to 76 in 1999." Finally, some targeted programs raise awareness about the conjunction of categories (e.g., geography and minority populations) with respect to diversity goals. For example, the goal to encourage and support underrepresented populations must also address geographic areas "with levels of poverty" or "urban sites [that] have developed limited research initiatives" (Urban Systemic Initiatives Program). There are geographic areas which remain resistant to initiatives to impact mathematics education at the elementary and middle school levels.

4. Does the process result in projects with a **positive societal impact**?

In the 13 CoV reports from **FY97**, the topic of *societal impact* was discussed primarily in terms of "advancing scientific progress" (Integrative Biology and Neurosciences) in general, and only secondarily in its application to societal or national needs. The fact that it may not be entirely clear what "effectively advancing the resolution of societal concerns" means is suggested by CoV reports that point out "better informing society of this finding remains a challenge for the future." That is, programs must take the initiative to explain to the general public *how* their research is of social benefit. As compared to CoV reports from FY99, however, it is somewhat surprising that concern with the *societal impact* of projects is relatively small.

The greatest impact on **FY99** CoV reports that discuss *societal impact* is the existence of Criterion 2 (and GPRA Goal #2). However, this does not entail that the meaning of *positive societal impact* is any more clearly understood than it was in FY97. For example, "The use of two criteria for merit review is relatively recent, so the CoV found it difficult to make a definitive assessment. Presently, it appears to be minimally effective. A large fraction of PIs and reviewers did not address criterion #2, as described in the Guide to Programs . . . Perhaps the articulation of criterion #2 as it is stated in the Proposal Review Form No. 3145-0060 and in the Grant Proposal Guide, ie., 'What are the broader impacts

of the proposed activity?' could contribute to misunderstanding by PIs and reviewers." (Neuroscience) Almost the same range of possible interpretations of the meaning of *societal impact* found in the FY99 project proposals can be found in the FY99 CoV reports. Further, as with other objectives of the new merit review criteria, there is a frequent call for data and tracking: "A systematic investigation of the connections between NSF-stimulated discoveries and their use in service to society will require long-term monitoring of scientific outcomes. The nature of basic research often makes it difficult to predict which discoveries will lead to important applications" (Neuroscience). Again, a number of CoV reports were "troubled by the fact that there is no clear definition of 'service to society,' leaving the achievement of this goal in some dispute" (International Programs).

An additional sense of *societal impact* in Criterion 2 (and GPRA Goal #2) is also raised in the Neuroscience CoV report. "One component of Criterion 2 (What are the broader impacts of the proposed activity?) is the contribution of the proposed activity to educational goals. The CoV believes that a weakness in the NSF proposal and review process is that the educational components are often ignored or presented in a cursory manner. PIs may feel that to adequately address the educational issues, they will have to sacrifice valuable space within the limited length of the proposal that could be more effectively utilized in presenting the scientific merit of the proposal. The fact that few proposal reviewers address the educational component of the proposal . . . shows that educational issues receive low priority in funding decisions . . . To encourage PIs to address Criterion 2 issues (in nontargeted proposals), we recommend that a new section should be created to address these issues." The connection between *broader contributions to society* and education is also seen in terms of projects which "provide foundations of new technologies and industrial practices . . . making pervasive contributions in training a scientifically educated and literate workforce" (Materials Research).

Consistent with the need to provide an operational definition for "societal relevance" expressed earlier, some CoV reports emphasize that the occurrence of *societal impact* may be a long-term or an indirect outcome of basic research: "it was striking that many projects whose goals could be characterized as pure science came to have important societal relevance. We urge that the NSF continue its strong advocacy and support for pure science in no small part because so many discoveries of importance of society have been the serendipitous outcome of such work rather than the product of more applied research." (Anthropological and Geographical Sciences) Or in another CoV: "The practical applications of this research often may not become apparent for decades" (Environmental Biology).

A number of CoV reports found that "the revised review criteria were considered to be an improvement from the previous sent of four criteria" (Civil and Mechanical Systems of Engineering). The customary manner of demonstrating *societal relevance* in CoV reports is to provide essentially subjective examples of projects with "benefits to society" (Astronomy). This is not to say the examples are without validity, but it points out that the perspective of "value to society" is naturally made from the context of each discipline.

The matter of responsibility for determining *societal relevance* is increasingly regarded as one held by scientific practitioners: "All scientists and agencies must communicate the results of their science to the public . . . about how it benefits our society, sparking the imagination of the young" (Astronomy). However, in terms of the review process itself, there were competing viewpoints: "The CoV was split in its opinion of who should carry the principle burden of explaining the societal relevance of the proposed research. Some thought that the burden should be placed on the PIs to describe the relevance of their work. However, others on the CoV felt that it was inappropriate to have PIs justify to their peers the relevance of the proposed research ("*of course* they think it is societally important – that is why they are in the field too"). (International Programs)

Not surprisingly, a frequent interpretation of *societal impact* is that in which it is assigned to projects which have obvious practical applications. "Outputs of DMR research have profound impacts on society. Contributions to new materials and processes are used by virtually all manufacturing industries, and have been crucial to computational and telecommunications, electronics, transportation, energy production, and medical instrumentation and materials." (Materials Research)

5. Does the process foster the **integration of research and education**?

In the 13 CoV reports from **FY97** the topic of *integration of research and education* was often discussed in terms of admittedly subjective impressions without specific supporting data. For example, the "CoV has the subjective impression that IBN's impact on the integration of research and education has been substantial, although rigorous performance indicators that would allow an objective evaluation are not yet available." (Integrative Biology and Neurosciences) Recommendations also include the need to track graduate students and postdoctoral fellows after their training "to see the impact of NSF funding on their scientific careers." Other discussions focused on programs specifically intended to fund PIs seeking to pursue excellence in both research and education. Not all of these programs were perceived as successful. For example, there were "significant differences . . . between what was intended by CAREER and the reality . . . uncertainty for proposers, reviewers, and program directors alike as to the goals of the program; inadequacy in the review process for fairly and expertly evaluating the educational components of the proposals . . . unless the POs are convinced that the CAREER program has merit, it is doomed to failure."" (Oceanography/Applied Ocean Sciences) Similarly, the POWRE "review process is not consistent." Other programs (e.g., the Advanced Technological Educational Program, which promotes improvement in technology at the national and regional level through curriculum development and undergraduate and secondary schools) fared better, with "processes used to solicit reviews, recommend, and document proposals actions [that] have integrity and are efficient." Programs of established disciplines, e.g. Physics, typically were commended for providing "continued focus on fundamental research and training of physics students."

The 13 **FY99** CoV reports shared with those of FY97 a concern for lack of data on which to evaluate the effectiveness of the goal of *integration of research and education.* For example, even with a cluster (Neuroscience) strongly committed to implementing

effective and innovative educational activities related to neuroscience, "because of the relative lack of outcome data documenting the effectiveness of the Neuroscience Cluster's educational programs, this conclusion is tentative." Special programs such as CAREER appear to have made little improvement. "None of the CoV felt that the CAREER program had achieve its objectives." (Anthropological and Geographical Sciences) There was also concern that "many of the educational initiatives of the NSF are not closely linked to the disciplines , , , [we] encourage much close cooperation between the educational wing of NSF and the science programs." Nevertheless, the disciplines did see participation in dedicated programs such as REU, RUI, POWRE, and CAREER as holding promise to develop a diverse and technological workforce. (Astronomy)

From the standpoint of proposal evaluation, it is not surprising that "reviewers placed much more emphasis on the technical merits and impact of the proposed work than on the educational . . . aspects." (Materials Research) Some CoVs claimed encouragement of the integration of research and education as strong, but without examples. (Genetics and Biochemistry of Gene Expression)

Given the limited sample of CoV reports, there is a general sense that while NSF initiatives have made some impact on mathematics education at the elementary and middle school level, the impact on urban sites where there has been a limited development of research initiatives is closer to being simply a "good effort" with very limited results. (Urban Systemic Initiatives Programs) However, the inherent appeal of other fields – for example, Environmental Biology – may be a contributing factor to success in "providing world-class, professional experiences in research and education."

6. Is the process **simpler?**

The **FY97** CoV reports are not making deliberately comparative judgments; therefore, elements pertaining to this question can only be noted in terms of general areas of concern and recommendations relevant to the review process. For example, the concern that reviewers are nearly all "research types" results in the perception that most reviewers are not well suited for reviewing both the educational and the research components of proposals (Oceanography/Applied Ocean Science). CoVs also express frequent concern with "consistency and uniformity" in panel reviews (Chemical and Transport Systems Division), and discomfort with how information in the jackets is organized – accessing information sometimes difficult because of "extensive supporting documentation." As we have seen elsewhere, there is also a general call for longer tracking of data and impact (Statewide Systemic Initiatives).

The review process in traditional disciplines such as Physics is generally felt to be "consistent and efficacious." In certain fields, the merging of processes for review of operations (e.g., the Oceanographic fleet) and proposals is a source of concern, with the recommendation that the processes be kept separate. Many CoVs note a lack of a sufficient number of ad hoc reviewers (Cell Biology), and a need to expand and improve the response rate.

FY99 CoV evaluations of the review process are generally more detailed, and assess process efficiency against the NSF goal of processing 79% of proposals within six months. CoVs are also sensitive to the qualitative workings of the process – e.g., panel reviewers having "thoughtful comments," "appropriate expertise," and evidencing a "great deal of agreement among the reviewers." (Neuroscience) CoVs also call for more ad hoc reviewers (e.g., at least five) and improving the return rate of 50%.

FY99 CoVs make an even stronger call for the need to collect data about diversity within reviewers (Anthropological and Geographical Sciences).

Many CoV reports note that for most reviewers, Criterion 1 carries more weight than Criterion 2. Depending on the scientific field or program, this is viewed as desirable or undesirable. The review process is generally felt consistent with the priorities and criteria stated in NSF solicitations, announcements, and guidelines. Some CoVs (Civil and Mechanical Systems of Engineering) found the use of the new NSF merit review criteria successful, with reviewers addressing both criteria, and "the revised criteria . . . considered to be an improvement from the previous set of four criteria." While review responses remain largely free-form, "this situation is somewhat improved with the two broader new criteria for FY98" (Materials Research). At the same time, in some fields the "use of both review criteria, scientific merit and broad impact, is occurring in less than 50% of the application evaluations." (Genetics and Biochemistry of Gene Expression) Several CoVs suggest requiring more specific instructions to use Criterion 2 on FastLane as well as the requirement that "applicants include a ½ - 1 page statement on the projected broader impact of the proposal." Where there is explicit discussion of FastLane, it is generally positive. However, "if the new Merit Review Criteria are to continue to be used, NSF needs to do a better job educating and coaching reviewers in their use." (International Programs)

For those disciplines favoring panel review, a commonly expressed supporting reason was that panels allow for "frank and open discussion of the proposal's strengths and weaknesses" (e.g., Upper Atmospheric Research Section)

FY99 CoVs often made specific recommendations on ways to improve the review process. For example, "Increased reviewer focus on criterion 2 might begin with the proposal – if proposals contain an explicit statement or section addressing criterion 2, ad hoc reviewers are more likely to use that criterion in their reviews." Again, "The review template should be modified by providing separate sections for addressing criterion 1 and 2 rather than listing both at the top of the form. The review form could also cite examples of broader scientific impact, such as graduate student training, public education, etc." (Environmental Biology) However, some of the suggestions carry implications that raise ethical questions and require further reflection: "Program officers and guidelines should encourage applicants to highlight and address clearly criterion 2 . . . citing the increased chances of a proposal's success if both criteria are addressed."

Discussion of Forms Used for CoV Reports and GPRA Questions

Background

For the FY 1997 Committee of Visitors reports, the only guidance provided by NSF to the committees was a general set of issues that the reports were required to address. In FY 1999, the CoV process was adapted to accommodate the GPRA format, and a specific list of questions was provided.

A memo from the GPRA Implementation Infrastructure Group to the Director's Policy Group dated August 17, 1999 discusses the call for FY 1999 Directorate GPRA Performance Reports. The memo reviews the GPRA Act of 1993 requirement that each federal agency provide a five-year Strategic Plan, annual Performance Plan, and an annual Performance Report. NSF submitted its first GPRA Strategic Plan to Congress in October 1997, and its first annual GPRA Performance Plan in March 1998. Its first Performance Report was presented in March 2000, providing a broad comparison of how NSF performed in comparison with the goals described in the 1999 Performance Plan.

The NSF GPRA Performance Report aggregates results across all programs, divisions, directorates, and offices. To accomplish this aggregation, standardized guidelines for CoVs and a standard reporting template for CoV reports were developed.

The guidelines for preparing FY 1999 Directorate GPRA Performance Reports includes an overall template, one part of which addresses the use of merit review criteria. This report – to be not more than one page – includes discussion of the use of merit review criteria by reviewers, and a discussion of the use of merit review criteria by program staff. Only a few of the CoV reports evaluated for the Academy study addressed one or both of these areas.

FY 1999 and FY 2000 Report Templates for CoVs

The report templates for FY 1999 and FY 2000 are similar; differences relevant to consideration of the merit review criteria are noted.

Section A of the report template addresses the **Integrity and Efficiency of the Program's Processes and Management.** Item 1 in this section asks CoVs to evaluate the *effectiveness of the program's use of merit review procedures.* This evaluation is to include the overall design and appropriateness of review mechanisms (panels, ad hoc reviews, site visits); the effectiveness of the review process; its efficiency and time to decision; the completeness of documentation making recommendations; and consistency of priorities and criteria stated in program solicitations, announcements, and guidelines. The template for FY 2000 adds that *constructive comments indicating areas for improvement are encouraged.*
Item 2 in this section asks CoVs to assess the *program's use of the new merit review criteria.* Here, the template for FY 1999 allows the following three evaluations:

- **Successful:** both review criteria are addressed by reviewers and used in the program officer's recommendation
- **Minimally effective:** reviewers use only one of the review criteria
- **Ineffective:** review criteria not used

The template for FY 2000 modifies this to allow the following evaluations:

> a. The program is successful when reviewers address the elements of both generic review criteria appropriate for the proposal at hand and when program officers take the information provided into account in their decisions on awards.
> b. Identify possible reasons for dissatisfaction with NSF's merit review system.

While (b) is admirable, (a) is rather confusing and appears to deflect from NSF's specific goal of requiring reviewers and Program Officers to use *both* criteria since it introduces several indeterminate qualifications.

Item 3 in section A asks CoVs to assess *reviewer selection*. This is to include the use of an adequate number for a balanced review; use of reviewers having appropriate expertise/qualifications; use of reviewers reflecting balance among characteristics such as geography, type of institution, and underrepresented groups; recognition and resolution of conflicts of interest. With respect to item 3, the Academy study notes several CoVs calling for a need for data about reviewer diversity so they can adequately evaluate these factors.

Item 4 in section A asks CoVs to assess the *resulting portfolio of awards* in a number of areas. Those relevant to the objectives of the new review criteria include effective identification of emerging opportunities; support of new investigators; support for integration of research and education; encouragement and support for participation of underrepresented groups.

Section B of the report templates address **Results: Outputs and Outcomes of NSF Investments** in direct response to GPRA Outcome Goals. GPRA outcomes 1, 2, and 3 are those most specifically relevant to the new merit review criteria. They are:

> 1. Discoveries at and across the frontiers of science and engineering that result from NSF investments.
> 2. Connections between discoveries and their use in service to society that result from NSF investments.
> 3. A diverse, globally-oriented workforce of scientists and engineers resulting from NSF investments.

First, it should be noted that a normal reading of *service to society* in Outcome 2 conveys a sense of broader "benefit to humanity" at least as much as it does of "utility" or "practical application." That is, it would be harder to restrict *societal impact* to mean only "utility" from this outcome goal. However, the template's elaboration of the intention of this goal does not reinforce this, or any, interpretation of the goal. The elaboration simply

says the program should be evaluated as successful "when the results of NSF awards are rapidly and readily available and feed, as appropriate, into education, policy development, or use by other federal agencies or the private sector."

As in Section A, the FY 2000 template replaces the categories of minimally effective vs. successful with a specification only of what successful is. The language of the category for successful, however, remains the same. The purpose of this shift in evaluation categories is not entirely clear.

Accompanying the report templates is a one-page introduction to the **Core Questions for CoVs**. The discussion is essentially the same for FY 1999 and FY 2000. NSF indicates that the judgments of CoVs are an important part of its continuous improvement process as well as necessary for compliance with the 1993 GPRA act. CoV reviews are expected to address both the *processes* leading to awards and the *results* of NSF investments.

CHAPTER 5: ANALYSIS OF INPUT FROM INTERVIEWS WITH NSF REVIEWERS

Summary of Findings

The following are major themes that emerged from interviews with ten NSF reviewers selected from a list of reviewers provided by NSF. At the end of this chapter is a reviewer survey form, originally developed for mail distribution. NSF preferred not to use a widely-distributed reviewer survey, but instead wished to have the interviews conducted more informally through telephone interviews. The questions in this form, therefore, were used as a general point of departure for the interview discussions, rather than as a fixed questionnaire. Also attached is the email letter from NSF to reviewers requesting their participation in the interviews.

- Most reviewers (80%) generally felt the new merit review criteria had made little or no contribution to achieving the several goals identified by NSF in instituting them. While some reviewers (20%) felt the goals were very desirable, many (roughly half) felt the language of Criterion 2 was vague, making the criterion hard to implement. Reviewers found some questions in Criterion 2 difficult to interpret. This was particularly true for the question regarding *benefits to society.* One reviewer indicated he "did not understand what NSF had in mind by *benefits to society* – there were too many different opinions about what it might mean, resulting in interpretations that were subjective." Many reviewers also felt the general goals had already been present in the old criteria, and in some cases were better expressed there. To use Criterion 2 properly and with consistency, therefore, reviewers strongly urged NSF provide examples of desired outcomes.

- A smaller percentage of reviewers (roughly 33%) were actually resistant to the goals of the new criteria. Some reviewers felt these goals were not applicable to the kinds of grant they reviewed (particularly those in traditional disciplines); other reviewers indicated they simply refused to apply Criterion 2 on the grounds that they did not find considerations of societal impact or infrastructure relevant or meaningful. One reviewer said they would downgrade a proposal if it lacked scientific merit but was only "trying to be relevant."

- For the reviewers who intended to apply both criteria, the most frequent procedure was to use Criterion 1 as a cut-off, looking at scientific merit first, then apply Criterion 2 to evaluate any remaining proposals. Reviewers who tried to apply Criterion 2 as a matter of course in their own evaluation process, generally found its language reasonably clear. However, even reviewers who tried to apply Criterion 2 felt it played a more minor role than Criterion 1. Therefore, it seems reasonable to infer that Criterion 2 was not being used in a balanced way or with equivalent weight to Criterion 1.

- To achieve the goals of the new criteria, most reviewers felt that creating targeted programs (rather than building the goals in some fashion into all proposals) represented the best strategy. Adequate guidance from NSF and clarity about desired outcomes would be a key to success in achieving those goals. Targeted programs to meet the objectives of providing benefit to society, and the other objectives of Criterion 2, should be kept separate.

- Reviewers generally felt that panels provided a more balanced review of proposals than individual review. However, it would be important for the Program Officer to choose a balanced panel and remove any bias, which can surface more easily in a panel.

- A number of reviewers liked the concept of FastLane but felt it needed more thought to avoid some areas of confusion. For example, sending graphics and maps was problematic.

- Some reviewers who also submitted proposals felt the process should allow proposers to respond to a review before the deadline. The concern was that a reviewer of limited expertise could kill a proposal for incorrect reasons.

- Reviewers by and large felt the review system was working fairly well. Most indicated that more time for review and more people doing reviews would improve it considerably.

- Reviewers who had read the recent NSF directives on the use of the new merit review criteria indicated that they had little impact on their use (or non-use) of Criterion 2.

- Some reviewers experienced greater difficulties in the use of the new review criteria as compared to the old. These difficulties often revolved around (the perception of) the "broad, abstract language" of the new criteria. The result was to give individual reviewers greater latitude and flexibility; at the same time, reviewers felt it also tended to make judgments more subjective and idiosyncratic. A small percentage of reviewers (20%) found the language of the new review criteria conceptually a bit clearer than the old because the new criteria established two distinct "chunks" rather than four considerations that sometimes overlapped.

- A number of reviewers indicated that NSF had to give better guidance and instructions to reviewers, including the specific mandate that reviewers address *both* criteria, assuming that to be an NSF goal. However, examples of where the concerns of Criterion 2 "were and were not relevant" should be included.

- Reviewers also expressed concern that the review process should better coordinate the roles and responsibilities of Program Officers vis à vis the responsibilities of reviewers. For example, it was suggested there should be more specific guidelines to ensure that reviewers are making judgments with sufficient consistency. There should

also be agreed upon criteria or guidelines for overriding reviewers' evaluations by Program Officers.

- Most reviewers interpreted the goal *to foster the integration of research and education* as identical with having an established program in graduate education in that scientific discipline.

NSF Reviewer Survey

(Note: This letter was not mailed to reviewers.)

To:	NSF Reviewers		4/20/00
From:	Dr. Robert R.N. Ross	Paul Herer	
	Academy Project Manager	NSF Office of Integrative Activities	
Subject	Reviewer Survey		

At the request of Congress, the National Science Foundation (NSF) has engaged the National Academy of Public Administration (the Academy) to conduct a study of the effectiveness its new merit review criteria for project selection.

As you are aware, the merit review process enables NSF to evaluate 30,000 proposals submitted to it annually, of which it funds approximately one third. In 1981, the National Science Board (NSB) adopted four criteria for the selection of research projects: (1) research performance competence, (2) intrinsic merit of the research, (3) utility or relevance of the research, and (4) effect of the research on the infrastructure of science and engineering. In May 1996, the NSB established an NSB-NSF Staff Task Force, charging it to re-examine the merit review criteria and make recommendations on retaining or changing them. In July 1997, NSF announced changes in its merit review criteria (Important Notice No. 121, *New Criteria for NSF Proposals*). The changes reflected its own analysis and input from the scientific and academic communities. The process resulted in the two criteria now in effect: (1) What is the intellectual merit of the proposed activity? and (2) What are the broader impacts of the proposed activity?

The enclosed survey seeks the valuable input of NSF reviewers about their experience in using the new merit review criteria.

The survey should take approximately 20 minutes to complete. Part I seeks general perceptions of the efficacy of the new criteria. Part II seeks specific suggestions about what is working and what isn't.

Please return the survey in the stamped return envelope to NSF at the address below no later than May 28. Please also feel free to contact Dr. Ross to discuss any aspect of this survey. If you would prefer to provide your response via telephone conversation or email, this can be arranged by contacting Dr. Ross at rrnross@vpm.com.

Thank you for your help and participation.

NSF Reviewer Survey
c/o Paul Herer
Office of Integrative Activities
The National Science Foundation
4201 Wilson Boulevard
Arlington, Virginia 22230

NSF Reviewer Survey

Part I. Your overall perceptions of the new merit review criteria

The following are six goals identified by NSF in using the new merit review criteria. Please indicate your perception of the relative contribution of the new criteria towards achieving each of these goals. (Low = little contribution; High = great contribution).

	Circle One				
	Low				**High**
1. Encourage a broader range of projects to be supported by NSF.	**1**	**2**	**3**	**4**	**5**
2. Seek wider institutional participation (e.g., by smaller as well as larger institutions).	**1**	**2**	**3**	**4**	**5**
3. Encourage greater diversity of participation in NSF funded projects by underrepresented minorities.	**1**	**2**	**3**	**4**	**5**
4. Support projects with positive social impact.	**1**	**2**	**3**	**4**	**5**
5. Foster the integration of research and education.	**1**	**2**	**3**	**4**	**5**
6. Simplify the merit review criteria.	**1**	**2**	**3**	**4**	**5**

Additional Comments

_____
_____
_____
_____
_____

Part II . Your specific analyses and recommendations

1. Describe any significant changes in your approach to making judgments about proposals using the new criteria (in comparison to how you used the old criteria).

_____
_____
_____
_____
_____
_____
_____

2a. List the top three difficulties you have experienced in using the new merit review criteria.

1_____
2_____
3_____

2b. List the top three advantages you have experienced in using the new merit review criteria.

1_____
2_____
3_____

3. Describe any significant differences you see in the use of the review criteria among the three modes of proposal review (mail review, panel review, combined mail and panel review).

_____
_____
_____
_____
_____

4. List the three areas of greatest impact on the review process from the use of the new merit review criteria (including recent directives for applying the new criteria more rigorously and systematically).

1_____
2_____
3_____

5a. List three ways in which the *language* of the new merit review criteria could be improved.

1_____
2_____
3_____

5b. List three ways in which the *process for using* the new merit review criteria could be improved.

1_____
2_____
3_____

E-mail letter from NSF to reviewers requesting participation in interviews

Dear Dr. _____

I am writing to ask for a few minutes of your valuable time to participate in a small, informal telephone survey of NSF reviewers about their experience in using the new merit review criteria.

At the request of Congress, the National Science Foundation (NSF) has engaged the National Academy of Public Administration (the Academy) to conduct a study of the effectiveness its new merit review criteria for project selection.

As you are aware, the merit review process enables NSF to evaluate 30,000 proposals submitted to it annually, of which it funds approximately one third. In 1981, the National Science Board (NSB) adopted four criteria for the selection of research projects: (1) research performance competence, (2) intrinsic merit of the research, (3) utility or relevance of the research, and (4) effect of the research on the infrastructure of science and engineering. In May 1996, the NSB established an NSB-NSF Staff Task Force, charging it to re-examine the merit review criteria and make recommendations on retaining or changing them.

In July 1997, NSF announced changes in its merit review criteria (Important Notice No. 121, New Criteria for NSF Proposals). The changes reflected its own analysis and input from the scientific and academic communities. The process resulted in the two criteria now in effect: (1) What is the intellectual merit of the proposed activity? and (2) What are the broader impacts of the proposed activity?

The telephone survey will be conducted by Dr. Bob Ross, a consultant to the Academy. It should take approximately 15-20 minutes to complete. Part I seeks general perceptions of the efficacy of the new criteria. Part II seeks specific suggestions about what is working and what isn't.

While we're hoping you will choose to participate in the survey, please feel free to decline to participate for any reason, by contacting Dr. Ross at rrnross@vpm.com, or when he contacts you by telephone.

Thank you for your help.

Paul Herer
Senior Staff Associate
Office of Integrative Activities
pherer@nsf.gov
703-306-1040

National Science Foundation
4201 Wilson Boulevard
Arlington, Virginia 22230

Discussion of Selection of NSF Reviewers

Chapter 2.2 briefly discusses the process and guidelines for the selection of reviewers. The following discusses that process in more detail – as it is relevant to the merit review criteria – on the basis of Sections 122.2 ff of the *Proposal and Award Manual* provided in ascii text by OIA.

The *Manual* indicates that peer review generally takes the form of ad hoc mail reviews, reviews by an assembled panel of experts, or a combination of both. Where appropriate, site visits are also employed. Each program has one primary method for peer review which represents the minimum evaluation received by proposals in that program. The primary method of peer review then may be supplemented with additional reviews or site visits.

NSF prefers that all proposals be reviewed by four to eight reviewers. When fewer than three written reviews or when a panel review involving fewer than three persons constitutes the external peer review, a justification must be given.

The NSF guidelines for selection of ad hoc reviewers are designed to ensure the selection of experts who can give Program Officers the proper information needed to make a recommendation in accordance with the criteria for selection of research projects. Since not all the criteria apply equally to every proposal or every Program, the balance among the criteria will influence the selection of reviewers. Program Officers also consider any specific criteria stated in program announcements and solicitations when selecting reviewers.

Reviewers should have special knowledge of the science and engineering subfields involved. This is intended to correspond to a balanced evaluation of proposals as related to competence, intrinsic merit, and utility of the research (some language of instructions to reviewers in the weighting of criteria notwithstanding). Reviewers should also have a broader or more generalized knowledge of the science and engineering subfields involved.

In addition, reviewers should have a broad knowledge of the infrastructure of the science and engineering enterprise and its educational activities. This relates to societal goals, scientific and engineering personnel, and distribution of resources to institutions and geographical areas. To the extent possible, reviewers should reflect a balance among various characteristics such as geography, type of institution, and underrepresented groups.

The *Manual* indicates that the guidelines for the selection of ad hoc reviewers apply to the choice of advisory committee/panel members as well, although it recognizes that it is seldom possible to meet every criterion mentioned above in a small group of people reviewing a variety of proposals. As above, the *Manual* emphasizes it is important that such groups be structured to provide broad representation and many views on matters under the advisory group's purview.

The *Manual* recommends general considerations that should help achieve reasonable balance in advisory groups, including the following:

- public impact—where pertinent, some members should be representative of regions, organizations, or segments of the public directly affected by issues under consideration.
- academic and nonacademic impact—members from the academic community should represent small, medium, and large institutions, as well as public and private institutions. (This is one of the few specific references to the new criteria's objective of *wider institutional participation*).
- under-represented views—special attention should be paid to obtaining qualified persons from underrepresented groups such as minorities, women, and the handicapped.
- age distribution—members should be selected from as broad a range of age as is feasible.
- geographic balance - members should be drawn from as broad a set of geographical areas as is feasible.

Regarding information for reviewers, the *Manual* states that a request to review a proposal should direct the reviewer to criteria for judgment and the relative importance of the respective criteria. No criteria may be included which were not described to the applicant. Letters to reviewers should include mention of the evaluation criteria printed on the back of the rating form, reference to the *Proposal Evaluation Form* (NSF Form 1), and a desired deadline for mailing the review (or for its receipt).

CHAPTER 6: ANALYSIS OF INPUT FROM NSF PERSONNEL, EXPERTS, AND STAKEHOLDERS

Summary of Findings

This chapter will capture the major themes and recommendations from a series of interviews, both formal and informal, held with NSF staff and management, external experts from several scientific disciplines in the research and academic communities, and other stakeholders. Many interviewees indicated they wished to remain anonymous. Therefore, this chapter will operate from the position that confidentiality will be maintained for *all* individuals providing input. Instead, comments and recommendations will be grouped under appropriate categories. Approximately 30 individuals were interviewed.

The interviews were wide ranging in subject matter, but the predominant themes that emerged were the following:

1. The motivations for instituting the new merit review criteria
2. The use of the merit review criteria in the review process
3. Ongoing issues in the merit review system
4. Evaluation of the merit review criteria
5. Improvements to the merit review process

Motivations for Instituting the New Merit Review Criteria

- This concern is behind the initiatives seeking proposals to now include integration of education and research. Even basic research proposals should now show some social impact. NSF also believes Senator Bond was responding to concerns of the scientific community in the change from the old to the new criteria: viz., to make sure no harm has been done in the change.

- Simplification was a major reason NSF chose to modify the review process by reducing the number of criterion from four to two.

- CoV reports from the late 1980s and early 1990s played a role in prompting the investigation into the merit review process.

- One of the biggest concerns leading to establishing the new criteria has been the need to broaden participation. The past system allowed the domination of big players, big universities. The updated merit review process is intended to provide a greater sense of fairness.

- The NSB had suggested a need for changes in the merit review criteria as early as 1995. A task force was put together to determine how the criteria were being used,

and to identify what were the most difficult aspects in using them. Then, focus groups were conducted at all levels – POs, administrative officers, all participants in the review process. They found that two of the four old criteria (the one on relevance and the one on infrastructure) were being ignored because people didn't know how to assess these factors. It was concluded that more focus to the merit review process was needed.

Use of the Merit Review Criteria

- The weight of each criterion depends on the nature of the proposal. The bulk of proposals are research-based and therefore evaluated primarily in terms of intellectual merit (Criterion 1). An educational proposal on the other hand would depend more heavily on Criterion 2.

- The language on the back of review forms is the primary guideline for reviewers on how to apply the criteria.

- NSF is concerned with developing and maintaining the scientific and engineering workforce. Therefore, proposals integrating research and education are rated highly.

- In actual practice, reviewers generally use Criterion 1 for their initial cut off. That is, if a proposal cannot first satisfy Criterion 1, it is not considered further. If it does satisfy Criterion 1, then Criterion 2 (social impact, etc) is used to choose among the remaining proposals. The exception is that specifically educational proposals might be evaluated using Criterion 2 first.

- Reviewers, in some instances, may give lip service to the use of Criterion 2. Program Officers have the final say and can override reviewer evaluations or apply Criterion 2 more.

Ongoing Issues in the Merit Review System

- Measuring diversity in the scientific community has remained difficult, because it is difficult to require people to furnish this information. The process is self-selective and there is no way of benchmarking or validating data. Outreach programs to increase diversity do not solve the measurement problem.

- Different divisions have different standards for how they approach reviews and grants. For example, information on ethnicity/race/gender of Principle Investigators is not gathered consistently across divisions.

- Historically there has been a tendency for a kind of "old boy network" to develop in the sense that the reviewers and proposers support one another in a closed system. New fields therefore become harder to award; new fields of research are also more prone to conflict of interest concerns because of the small numbers of researchers in these fields.

- A problem with the current proposal forms is that not enough attention is drawn to both criteria. Clear language and guidelines for good proposal writing are needed.

- Reviewers are not compensated but volunteer their time. This often makes it difficult to get reviewers.

- NSF has not communicated adequately to the scientific communities what it means by "integrated research and education" (i.e., over and above the tradition of using graduate students in research projects). NSF needs to provide more concrete examples for how to support Criterion 2 using cases that visibly show what goals are desired, not more rules.

Evaluation of the Merit Review Criteria

- NSF management generally believes it is not possible at this point to get a statistically robust evaluation to determine if any differences result from the new criteria. However, NSF is aware of the minimal use of Criterion 2 by reviewers.

- Experts outside of NSF also feel it is not possible to fully determine the impact of the new criteria because the time of their being in use has been too short, and one can't isolate all the variables. Many experts also feel that Criterion 2 is always in potential conflict with "best science." These individuals believe it might be better to set aside money for specific social objectives and develop more institutions.

- NSF management is aware most people have felt that going from the old four criteria to the new two criteria has made no real change. The new Criterion 2 now just forces some insistence to *write* to the issues of Criterion 2, whereas previous discussions tended to be only about what applied to the proposal.

- NSF has been moving in the direction of Criterion 2 for five or six years. This is quite natural, since NSF was chartered to have responsibility for maintaining the scientific workforce, and this automatically drives a responsibility to consider the social impact of research projects and the diversity concerns of Criterion 2.

- Some scientific communities have found Criterion 2 hard to accept. NSF received approximately 300-400 emails on the new criteria that showed a strong bifurcation of opinion. Approximately half saw NSF as having been too elitist and therefore welcomed the change to the new criteria. Half remained purists and didn't like the new criteria. Mathematicians, for example, were against the new criteria. Geophysicists have been for them. NSF then backed off and said reviewers did not need to apply the criteria equally; reviewers needed to apply Criterion 2 but *only in some degree*. This effectively made use of the new criteria more similar to the old criteria, whose language gave reviewers considerable freedom and allowed them to apply the criteria differently. When NSF saw Criterion 2 being ignored, a GPRA

performance goal was defined for it, although this has not yet been fully implemented.

- While examining proposal jackets may give some sense of to what extent reviewers are paying attention to the new criteria, many experts believe the proposal jackets will not reveal much about changes in what NSF is funding.

- There was no electronic submission process for the old merit review criteria; all was done on written forms. Therefore, two confusing things occurred at once in 1998. Many people saw the old and new criteria as essentially the same.

- Earlier CoV reports had no template, no structure. There were guidelines in the Program Administrator's Manual (PAM), but they were not followed. The older CoVs asked different questions than current ones, and therefore displayed greater variance. The newer CoVs are more uniform in approach and more reflective, whereas the older ones were mostly just a report.

Improvements to the Merit Review Process

- NSF has identified 16 options to improve the merit system, and improve proposers, reviewers, and program manager's use of the review process. The SMIG (Senior Management Integration Group) gave the go ahead to implement 11 of the outlined options. One option uses the electronic system to remind reviewers to address both criteria.

- The current major evaluation of the review process consists of qualitative assessments by CoVs. NSF has made moves to broaden participation within CoVs and include more industry players.

- There is a need for NSF to remind people to use Criterion 2. To push this, NSF will institute a number of mechanisms to force reviewers to discuss Criterion 2, and will mandate that Project Officers comment on Criterion 2. NSF will also conduct training sessions for division directors.

- The NSB is ultimately responsible for approving new criteria. Changes in instructions to reviewers to implement Criterion 2 will make it clearer that reviewers must provide separate text for Criterion 1 and 2, rather than merge them.

- Some differences have occurred with the new electronic submissions: proposals and reviews are more terse, more directed, with somewhat greater sensitivity to issues, and a greater awareness of NSF policies from interacting with the website.

- NSF management will use GPRA to drive the goals of the new criteria. The goals of new criteria did antecede GPRA but now play together. NSF supports GPRA because, among other things, it helps justify budget increases. Program Officers generally think GPRA is useful.

- NSF will emphasize the importance of better implementation of the new merit review criteria through new directives, letters to institutions, and letters to proposers and reviewers.

Discussion of NSF and Comparative Assessments of the Review Process

Summary of Findings

- In 1999 NSF's American Customer Satisfaction Index (ACSI) of 57 was 15 points below the current national ACSI of 72 and 8 points below the current government (IRS and USPS) sector index of 65. NSF's interpretation of these results would appear to be a matter for further discussion and analysis. NSF interprets these results to indicate that applicants generally hold NSF in high regard and give the Foundation high marks for the accessibility and usefulness of its information, but only mid-level evaluations for the proposal review process and for its handling of customer complaints. It should be noted that NSF was one of the few agencies that included declinees in its pool of customers that were surveyed.

- NSF should accompany its plan to conduct additional surveys and identify effective practices for responding to customer complaints with a systematic root-cause analysis of those customer complaints before jumping to solutions.

- In its summary of FY 1999 GPRA performance results, NSF notes that areas requiring improvement include a need to show increases in participation of underrepresented groups in science and engineering, and the need to improve use of the new merit review criteria. A major weakness found by Committee of Visitors and Advisory Committee reports was that reviewers and applicants were not fully addressing *both* review criteria.

- Results from NSF Task Group studies indicate considerable variation in the use of the criteria across the several divisions of NSF. This applies to the old review criteria (where people found Criteria 3 and 4 difficult to apply and often ignored) as well as the new criteria.

- NSF Task Group reviewer surveys (including the 1991 survey of 9000 reviewers, *The Track Record of NSF Proposal Review, Reviewers Rate the Process*) indicated a lack of clarity and applicability especially in the old Criteria 3 and 4, and a sizable number of reviewers who simply ignored the criteria.

- The draft criteria components proposed by the 1996 NSF Task Group did not consider many of the objectives of the current merit review criteria.

- Comparative Assessments of the Review Process

In the NSF FY 2000 - 2005 Strategic Plan, among the implementation strategies under "Critical Success Factor 1" (operating a credible, efficient merit review system), it is stated that NSF regularly assesses performance of all aspects of the merit review system, comparing its efficiency, effectiveness, customer satisfaction and integrity against similar processes run by other organizations. In response to a request for documents from NSF of such comparisons of the merit review systems with other organizations, NSF provided the following:

- Government-Wide Survey of Customer Satisfaction, NSF 3 page draft dated December 1999
- Untitled 1 page draft dated 11/2/99 [brief summary of above draft]
- The Track Record of NSF Proposal Review: Reviewers Rate the Process. NSF Program Evaluation Staff and Science Resources International (SRI 1991) (fragments)
- Peer Review.: Reforms Needed to Ensure Fairness in Federal Agency Grant Selection, United States General Accounting Office June, 1994 (GAO/PEMD-94-1) and accompanying article in Washington Post, 7/28/94
- NSF Proposal Review Project Reports from 1996, including:
- Review Criteria : Report of Task Group 1h (P. Stephens, Chair), dated 2/14/96
- Interdisciplinary Proposals: Report of Task Group 1f (M. Burka, Chair), dated 2/13/96
- NSB Review Criteria: Options Discussion Paper, dated 4/10/96
- "Grants Peer Review in Theory and Practice", Daryl E. Chubin, NSF, *Evaluation Review*, Vol. 18 No. 1, February 1994, pp. 20-30
- Proposal Evaluation within other Federal Agencies, undated 2 page draft
- Task Group on Review Variations (D. Schindel/D. Chubin): listed but not provided
- Task Group on Calibration and Disaggregated Ratings (C. Eavey): not provided

In addition, several other documents were obtained directly:

- Observations on the National Science Foundation's Performance Plan for Fiscal Year 2000, Victor Rezendes, General Accounting Office, July 20, 1999, GAO/RCED-99-206R
- GPRA Performance Report FY 1999 (NSF 00-64)
- Observations on the National Science Foundation's Annual Performance Plan [for FY 1999], Susan Kladiva, General Accounting Office, May 19, 1998, GAO/RCED-98-192R
- Results Act: Observations on the National Science Foundation's Draft Strategic Plan, Victor Rezendes, July 11, 1997, GAO/RCED-97-203R
- Federally Funded Research: Decisions for a Decade, U.S. Congress, Office of Technology Assessment, May 1991 OTA-SET-490

Given the mixture of documents and time frames, their discussion will be ordered on the basis of significant new information or opinions beyond those already expressed in this report.

Government-Wide Survey of Customer Satisfaction

In 1999 NSF volunteered to participate with 30 other agencies in a national assessment of customer satisfaction sponsored by the President's Management Council and the Vice President's National Partnership for Reinventing Government. The stated goals of the process were:

- to set a baseline for measuring customer satisfaction in the federal government
- to benchmark customer satisfaction among federal agencies and the private sector
- to identify areas in which agencies can improve customer satisfaction

NSF's view was that the survey would provide useful information about the impact of its methods and processes on the scientists, engineers, and educators who apply for NSF grants.

The means for conducting this assessment was the American Customer Satisfaction Index (ACSI) – a cross-industry index of customer satisfaction established in 1994. The ACSI measures customer satisfaction for 170 private sector corporations and two major federal agencies (IRS and USPS). The National Quality Research Center at the University of Michigan Business School administered the assessment of customer satisfaction for NSF and the other federal agencies.

NSF chose all grant applicants (both awardees and declinees) in FY 1998 as the customer segment for which to measure satisfaction. NSF provided a random sample of 3000 names from the pool of approximately 28,000 grant applicants for FY 1998. The ACSI survey team took a smaller sample of 260 to interview. 68% of the applicants interviewed during the ACSI process submitted proposals which were declined by NSF. This was consistent with NSF's overall proposal success rate.

The following table summarizes NSF's results for key measurements resulting from the interviews of 260 FY 1998 grant applicants (all scores are on a 0-100 scale):

Measurement	**Description**	**Score**
Overall customer satisfaction	This measure is derived from responses to 3 general questions about overall satisfaction, expectations, and comparison to an ideal organization.	57
Expected/perceived quality received from NSF	These scores measure the anticipated quality of service and the experienced quality of service for NSF customers.	71/67
Applicant trust	The applicant trust index is derived from responses to 2 questions about confidence in NSF's ability to administer the process fairly and competently in the future and willingness to rely on NSF to do a good job of advancing scientific knowledge in the future.	67
Accessibility and usefulness of NSF	Measures customer ratings for the accessiblity of needed information and the usefulness (current,	80

information	accurate, helpful, and relevant) of the information provided.	
Timeliness and efficiency of the proposal process	This measure is based on a single question about the timeliness and efficiency of the proposal process.	56
Quality and fairness of merit review	This measure is based on two questions about the quality of review of the applicant's proposal and the fairness of the review and decision process.	58

In addition, the survey determined that 36% of those interviewed had ever complained to NSF in some way. The 36% of applicants who complained averaged 2.1 formal complaints, either in writing or by telephone, and 2.3 informal complaints while talking to personnel of NSF. Complainants gave NSF a score of 57 (on a scale of 0-100) for how well or poorly the most recent complaint was handled.

The NSF customer satisfaction index (ACSI) of 57 was 15 points below the current national ACSI of 72 and 8 points below the current government (IRS and USPS) sector index of 65.

NSF feels these results indicate that NSF grant applicants generally hold NSF in high regard and give the Foundation high marks for the accessibility and usefulness of its information. However, NSF received only mid-level evaluations for the proposal review process and for its handling of customer complaints. NSF felt that the fact that two thirds of the applicants surveyed were turned down was a contributing factor in the survey results.

During FY 2000, NSF indicated the following plans to respond to the results of the government-wide survey:

- Conduct additional surveys of applicants in the coming year to confirm the results of the ACSI and to get more detailed information on specific issues related to proposal review and customer interaction. The results of these surveys will help focus efforts to improve service in these areas.

- Identify effective practices for responding to customer complaints, both within NSF and in other organizations with similar customer interactions. NSF will disseminate information about these effective practices throughout the agency, identify promising models for customer service systems both inside and outside NSF, and pilot the best of these models in NSF divisions.

- Establish a GPRA annual performance goal for customer service that will use the ACSI as a key indicator. Many of current GPRA performance goals set performance measures for NSF customer service standards. Establishing this new goal will further increase awareness of customer concerns throughout NSF and set a pattern for continuous improvement in service to customers.

NSF GPRA Performance Report

The FY 1999 NSF GPRA Performance Report represents an assessment based on the first full year of GPRA implementation. A number of issues connected to the GPRA report have been discussed elsewhere in this study. The following focuses specifically on GPRA performance goals relevant to merit review – goals 6 and 7.

The performance results for GPRA Goal 6 simply assesses the *percentage of use* of merit review. This goal is stated in the standard GPRA (quantitative) format. Goal 6 states that "at least 90 percent of NSF funds will be allocated to projects reviewed by appropriate peers external to NSF and selected through a merit-based competitive process." NSF determined that for FY 1999 *this goal was achieved*. With a FY 1999 goal of 90%, in FY 1999 NSF indicates that 95% of projects allocated funds were merit reviewed.
The performance results for GPRA Goal 7 assesses the *implementation of the new merit review criteria*. This goal is stated in the alternative GPRA format. The *alternative* format is a qualitative scale that allows NSF to report whether it has been "successful" or "minimally effective" in achieving its outcome goals for those goals that do not lend themselves to quantitative reporting. Goal 7 states that "NSF performance in implementation of the new merit review criteria is successful when reviewers address the elements of both generic review criteria appropriate to the proposal at hand and when program officers take the information provided into account in their decisions on awards; minimally effective when reviewers consistently use only a few of the suggested elements of the generic review criteria although others might be applicable." Beyond this statement which contains such relatively loose expressions as "take the information provided into account" and "use only a few of the suggested elements", there is no further specific scale to the FY 1999 goal. NSF characterizes its FY 1999 results as *largely successful, needs improvement*. In FY 1999 a total of 44 reports by external experts (i.e., 38 Committee of Visitors reports and 6 Advisory Committees reports) rated NSF on their use of the new merit review criteria. NSF indicates that in FY 1999, external experts evaluated approximately 40% of NSF's 200 programs in (that is, 80 programs). Therefore, this suggests that these 44 reports represent an evaluation of GPRA Goal 7 by only slightly over half the external evaluations. Of the total of 44 reports by external experts rating NSF on their use of the new merit review criteria, NSF was rated successful in achieving this goal by 36 of those reports.

In its summary of FY 1999 GPRA performance results, NSF notes areas needing improvement as including the need to show increases in participation of underrepresented groups in science and engineering, and the need to improve use of the new merit review criteria. A major weakness found by CoV and AC reports was that reviewers and applicants were not fully addressing both review criteria.

NSF Task Group Studies of the Review Criteria

•

The Stephens' chaired "Review Criteria Task Group 1h" found that the weight assigned to the four NSB criteria by reviewers and Program Officers in the evaluation process varied considerably. Moreover, it was substantive comments on the technical aspects of

the proposals that played the most important role in the decision process. A 1991 NSF survey of 9000 reviewers (*The Track Record of NSF Proposal Review, Reviewers Rate the Process*) revealed that reviewers considered the first two criteria (intrinsic merit and PI competence) the most important. Views of Criterion 3 (utility or relevance) ranged from not applicable to very important. Reviewers found Criterion 4 (infrastructure) ambiguous and difficult to evaluate. Less than 50% of reviewers said they commented on all four criteria; 20% admitted to ignoring the criteria. Program Officers indicated that most reviewers did not comment on Criterion 4.

The Task Group's survey of NSF divisions also indicated that some programs employed special criteria in addition to the four generic criteria. These additional criteria were given at least the same or more weight than the generic criteria.

The Task Group noted that Criterion 4 (infrastructure) was found to be extremely broad and presented "a challenge in terms of interpretation and application." Reviewers frequently "lacked the basis on which to make a judgment"; in addition, program needs varied considerably with respect to this criterion.

There was division within the Task Group concerning the degree to which the four generic criteria should be revised. However, its primary recommendations were that:

- The NSB criteria are in need of clarification and should be rewritten, with consideration to (a) making the criteria clear to evaluators, (b) emphasizing important attributes such as innovation, clarity, soundness of approach, (c) encouraging substantive comments.
- NSF should explore more effective ways to apply the infrastructure criterion.
- NSF should continue to allow programs to employ additional criteria.

The Task Group proposed the following components to be considered in review criteria:

- Intrinsic Merit
- Significance
- Innovation
- Approach (technical soundness)
- Feasibility
- Effect of the project on the infrastructure of science and engineering

It should be noted that these proposed components of the merit review criteria do not address many of the objectives of the current merit review criteria.

The Task Group 1h also examined the relationship of the review criteria to the NSF Strategic Plan, and concluded that maintaining flexibility in how and which criteria are applied was the most effective way to make decisions consistent with strategic goals – in particular, identifying cutting edge projects that push back frontiers.

An NSB Review Criteria Options Discussion Paper (4/10/96) listed several functions review criteria are intended to serve: establishing standards to judge the quality of proposals; characterizing traits common to all projects supported by NSF; communicating goals and values to external communities. It can be noted that these functions also do not consider many of the specific objectives of the current merit review criteria.

This discussion paper presented four options related to the 1981 NSB review criteria:

1. Leave the criteria unchanged
2. Clarify and revise the criteria
3. Develop new criteria
4. Replace generic criteria with program-specific criteria

The paper did not recommend any one course of action; however, several interesting points to stimulate discussion emerged:

- The history, tradition, and acceptance of the 1981 generic criteria was a strong motivation to leave them unchanged
- On the other hand, lack of clarity in language could encourage the use of "unwritten" criteria (pointed out in the 1994 GAO report) as well as non-uniform application of the criteria by reviewers
- Perhaps the strongest motivation for developing a new set of criteria (in addition to a perception of lack of clarity in the language of the old criteria) was the belief that the old criteria were not applicable to some programs initiated more recently in NSF, in EHR and other areas (cf, Deputy Directory Memorandum of 9 June 1995)
- The advantages of moving towards program-specific criteria were outweighed by the worry of confusing PIs through a proliferation of specific criteria without any universal or connecting themes

The issue of the consequences of including selection criteria other than or in addition to the generic criteria raises a number of interesting issues, perhaps the most important of which is that of *fairness*. In his article "Grants Peer Review in Theory and Practice", Daryl Chubin of NSF points out that "peer review should . . . be *fair*, adhering to societal norms of equitable treatment as well as scientific norms of universalism and disinterestedness." However, "with the inclusion of selection criteria other than technical merit in peer review, fairness is often seen as diluting quality." It could also be argued that the existence of non-universally applied criteria also challenges the concept of *fairness* itself, since it alters the distribution of funding opportunities from a finite funding set.

The issue of *fairness* is also briefly discussed in the Office of Technology Assessment's *Federally Funded Research: Decisions for a Decade* (OTA-SET-490, May 1991, p. 129). OTA points out that "in addition to the mainstream disbursal of funds, agencies often allocate funds using other types of programs."

> "*Set-aside* programs are agencywide discretionary actions. They select one characteristic that captures a need not served by mainstream proposal review and restricts competition for research funding to a pool of eligibles who qualify by virtue of that characteristic. Thus, there are set-asides for women, ethnic minorities, young investigators, investigators located at traditionally nonresearch institutions, and investigators residing in States that have been underrepresented in the amount of Federal research funds they receive relative to their share of the general population or the number of undergraduates they enroll."

OTA points out that the assumption underlying set-aside programs is that there are capable researchers everywhere who – for lack of opportunity or obvious disparities in experience – are disadvantaged in the ordinary competitive proposal process. The solution is a separate competition, still organized around the criterion of technical merit, that pits like against like. NSF uses this to both develop institutional research capabilities and widen geographic diversity.

Proposal Evaluation within other Federal Agencies

An untitled, undated 2 page paper briefly sketches the major elements of evaluation criteria used by NIH, DARPA, TRP Development, and NASA. An additional set of undated overhead slides titled "Peer Review at NIH" suggests the use of criteria very similar to the NSF 1981 generic criteria, with the addition of one specifically directed to evaluate the "adequacy of plans to include both genders and minorities and their subgroups as appropriate for the scientific goals of the research." NIH was the original site of peer review in the Federal Government, beginning with the National Advisory Cancer Council in 1937 (OTA-SET-490, p 126). NIH has instituted a process that results in the establishment of scientific review groups of 18-20 members, serving multi-year terms. This suggests not only a way to ensure a "match of scientific content of application with the expertise of specific reviewers", but also the development of a more systematic, collegial approach to maintain an ongoing reviewer community. Among the factors reviewers evaluate are the inclusion of women, minorities, and children, and issues relating to the protection of human subjects, the environment, and animal welfare.

Peer Review: Reforms Needed to Ensure Fairness in Federal Agency Grant Selection, June 1994 (GAO/PEMD-94-1) examined grant selection by peer review in the National Institutes of Health (NIH), the National Endowment for the Humanities (NEH), and NSF. As indicated earlier, this study emerged from a long history of controversy about how peer review was practiced – in particular, whether existing systems were providing fair, impartial reviews of proposals. GAO collected files on a sample of grant proposals, approximately half of which had been funded. GAO also reviewed agency policy documents, and surveyed 1400 reviewers. All the study found that peer review processes were working reasonably well, agencies needed to "take a number of measures to better ensure fairness." Among particular areas of concern with NSF, GAO found that junior scholars and women were consistently underrepresented, and that there were problems in the consistency in how review criteria were applied. With respect to the latter, the study

found that reviewers often "used unwritten decision rules in rating proposals." The most common of these rules concerned the quality of preliminary work results.

GAO recommended that the Director of NSF (1) use targeted outreach efforts to attract young reviewers, (2) increase monitoring of discrimination in scoring, (3) employ a scoring system in which proposals are rated separately on a number of criteria as well as overall, and (4) inform applicants of any unwritten decision rules used by reviewers.

GAO Observations on NSF Performance Plans

The following summarizes the major relevant observations by GAO on NSF Performance Plans for FY 2000, FY 1999, and its draft Strategic Plan of 1997.

For the Annual Performance Plan of FY 2000, general weaknesses noted include the lack of "clear linkages between the budget and performance goals" and "limited confidence in the validation and verification of data." The latter concern is of particular interest to this study. GAO notes that the Performance Plan indicates moderate progress in addressing weaknesses identified in their assessment of the FY 1999 Performance Plan. However, "while NSF's performance plan provides a general picture of intended performance across the agency, there are still inconsistencies in the information supporting each performance goal."

NSF expresses annual performance goals in terms of "successful" and "minimally effective" performance, an alternative format allowed by GPRA and OMB.

> "For example, NSF believes its performance will be rated successful in meeting its strategic goal of promoting connections between discoveries and their use in service to society if the results of NSF awards are rapidly and readily available and, as appropriate, feed into education, policy development, or work of other federal agencies or the private sector.

But NSF's performance will be rated only minimally effective if the results of its grant awards show only the potential for use in service to society."

GAO notes that "the descriptive statements developed by NSF reasonably define the type and level of annual performance that the agency expects for these activities", and allows for expert judgement based both qualitative and quantitative information about performance. GAO appears generally positive about the "Guidelines for Committees of Visitors" which provide information on how various phrases relating to the evaluation of qualitative information are to be interpreted and used by reviewers. However, it also notes that variation in interpretations "may lead to difficulties in evaluating results over time." Further, "until the NSF guidelines for evaluating performance results are implemented, it will be difficult to assess whether they are providing enough guidance to aid reviewers."
GAO finds NSF's strategies to achieve its five primary goals for scientific research and education – including its use of a competitive merit-based review process – reasonable

approaches overall. However, it also feels "the agency's performance plan provides limited confidence that agency performance information will be credible." Largely, this is a complaint about the lack of standards or procedures that will be used to assess the reliability of four information systems that store, process, analyze, and report performance measurement data. (The 1999 GAO assessment expresses a somewhat broader criticism of the measures NSF uses to determine whether the agency is meeting the minimum or successful levels of performance in scientific research and education; see p. 11 ff.) It is not clear to what extent this specifically impacts the reliability of data for the merit review process, although concerns about the tracking of data relevant to the new Criterion 2 have been expressed elsewhere in this report.

In a table (appended to the GAO report) identifying management challenges confronting NSF as identified by its Office of Inspector General (OIG), the first item listed as an area of concern is "managing an effective merit review system." NSF's two goals related to merit review are (1) that at least 90% of NSF funds will be allocated through a merit-based competitive process, and (2) that NSF's performance in implementation of the new merit review criteria well be successful when reviewers address the elements of both generic review criteria appropriate to the proposal at hand and when program officers take the information provided into account in their decisions on awards; performance will be minimally successful when reviewers consistently use only a few of the suggested elements of the generic review criteria, although others might be applicable.

Clearly, to determine whether NSF has achieved "successful" or "minimally successful" performance on these standards for the merit review system will require tracking more than simply anecdotal or qualitative information as provided by Committee of Visitors reports. It will require quantitative tracking as well as qualitative. For this very reason, the Academy study of the merit review system cannot determine with any degree of certitude whether NSF's use of its new merit review criteria has, to this point, been "successful" or "minimally successful." The highly preliminary indication from interviews with reviewers, however, would suggest that – on the basis of achieving Goal 2 above – performance in implementing the new merit review criteria has at best been only "minimally successful." This is largely because, from the reviewers interviewed, Criterion 2 is either ignored or not taken seriously in evaluating proposals.

Bibliography