United States General Accounting Office

GAO

Report to the Chairman, Committee on Governmental Affairs, U.S. Senate

June 1994

PEER REVIEW

Reforms Needed to Ensure Fairness in Federal Agency Grant Selection

GAO/PEMD-94-1

GAO
United States
General Accounting Office
Washington, D.C. 20548

Program Evaluation and
Methodology Division

B-254742

June 24, 1994

The Honorable John Glenn
Chairman, Committee on
Governmental Affairs
United States Senate

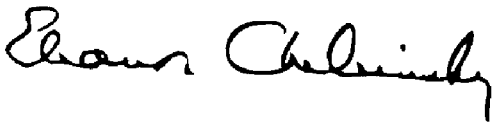
Dear Mr. Chairman:

Federal agencies throughout the government use peer review to evaluate research and other projects proposed for federal funding. Concerned about how peer review is working, you asked us to review its fairness. In this report, we examine peer review processes at selected federal agencies, concentrating on issues related to fairness in the selection of reviewers, the conduct of reviews, and the decision to award funds.

As agreed with your office, we plan no further distribution of this report until 30 days from its date of issue, unless you publicly announce its contents earlier. We will then send copies to the Director of the National Institutes of Health, the Director of the National Science Foundation, and the Chairman of the National Endowment for the Humanities. We will also make copies available to others upon request.

If you have any questions or would like additional information, please call me at (202) 512-2900, or Robert L. York, Director of Program Evaluation in Human Services Areas, at (202) 512-5885. Other major contributors to this report are listed in appendix VI.

Sincerely yours,



Eleanor Chelimsky
Assistant Comptroller General

Executive Summary

Purpose

Peer review is used throughout the federal government to evaluate research and other projects proposed for federal funding. The Senate Committee on Governmental Affairs asked GAO to examine the fairness of peer review processes in federal agencies. GAO first identified potential or perceived weaknesses in this area along with agency efforts to deal with them (appendix I) and then developed empirical evidence of the extent to which some of these weaknesses actually occur.

Background

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. GAO focused on the extent to which fairness problems occur in three areas: the selection of peer reviewers, the scoring of proposals by reviewers, and the final funding decisions of agencies.

GAO examined grant selection in three federal agencies that use peer review: the National Institutes of Health (NIH), the National Science Foundation (NSF), and the National Endowment for the Humanities (NEH). At each agency, GAO collected administrative files on a sample of grant proposals, approximately half of which had been funded. GAO then surveyed almost 1,400 reviewers of these proposals to obtain information not available from the agencies. In addition, GAO interviewed agency officials and reviewed documents to obtain procedural and policy information. GAO also observed panel meetings at each agency.

Results in Brief

Overall, peer review processes appear to be working reasonably well and are generally supported by peer reviewers. However, the agencies need to take a number of measures to better ensure fairness in the three areas of the study's focus. For example, with regard to reviewer selection, junior scholars were consistently underrepresented among reviewers at all three agencies. Further, in some NSF programs, women were also underrepresented. And although most reviewers reported expertise in the general areas of the proposals they reviewed, many were not expert on closely related questions, especially at NIH.

With respect to the rating of proposals, GAO found that much of the variation in peer review scores was unrelated to any measured characteristics of reviewers or applicants. This suggests that the intrinsic qualities of a proposal (such as the research design and the importance of the questions it addressed) were important factors in reviewers' scoring.

Nonetheless, scores were better for men than women at all three agencies and for whites than minorities at NSF. Data on the race of applicants were unavailable from NIH and NEH.

GAO also noted some problems in how review criteria were applied. Reviewers were inconsistent in considering agency criteria and, especially at NSF and NEH, in applying agency criteria and rating scales. In addition, across all three agencies, reviewers used unwritten decision rules in rating proposals; the most common of these rules concerned the quality of preliminary work results.

Finally, with regard to the relationship of peer review to funding decisions, the agencies varied in the extent to which peer review scores were in fact decisive in determining which proposals were approved for funding.

Principal Findings

Selection of Peer Reviewers

Reviewers were broadly similar to applicants on a number of dimensions. GAO found that, contrary to what some critics have asserted, reviewers were not more likely to come from elite institutions than were applicants, and there were few differences in region of origin. However, in some NSF programs, women were underrepresented among mail reviewers (those not present at panel meetings); this is important because some programs rely heavily, or exclusively, on mail reviewers.

At all three agencies, large majorities of reviewers (from 66 percent at NIH to 93 percent at NSF) reported that their own work was at least in the general area of the proposals they reviewed. But only a minority said their work was on the same or related questions (from 14 percent at NIH to 44 percent at NSF); and only half or less said they could cite much of the literature (from 21 percent at NIH to 51 percent at NSF). However, GAO found a tradeoff between such expertise and the likelihood that reviewers personally knew applicants, a potential source of bias. Thus, NSF reviewers generally showed the most relevant expertise but were also most likely to personally know the applicants (46 percent, compared to 23 percent at NIH); NIH reviewers were less likely to report both directly related expertise and personal knowledge of the applicant.

Factors Related to Scoring Proposals

Several factors alleged to affect reviews were not, in fact, related to scores at any of the agencies: the reviewer's proximity of scholarly interest or knowledge of the literature and the applicant's region, academic rank, or even employing department's prestige.

However, a number of other factors, in addition to gender (and race at NSF), were related to the scores given to proposals. At all three agencies, reviewers tended to give better scores to proposals from applicants perceived to have stronger publication track records, a measure that is widely used as an indicator of an applicant's demonstrated potential to successfully complete a project. In addition, at both NSF and NEH, reviewers gave better scores to applicants whom they knew than to those they did not know. Logically possible explanations for these results might include that (1) experienced, well-known, white, male scholars write better proposals than others or (2) they know the rules and norms for proposal-writing better or (3) some bias in the scoring of proposals exists at these agencies.

Factors Related to Decisions on Funding

The influence of different factors on final decisions regarding funding varied by agency. At NIH, the review panel's score was the only factor that was significantly related to whether the grant was awarded funding. At NSF, however, getting a good score was more important for little-known scholars than for those who were well known; this difference became more important the higher the amount requested. Finally, at NEH, beyond some point, the odds of funding decreased sharply for proposals with worse scores and higher requested amounts.

Recommendations

GAO recommends that the Director of NIH, the Director of NSF, and the Chairman of NEH (1) use targeted outreach efforts, at least experimentally, to attract young reviewers; (2) increase the monitoring of discrimination in scoring, including conducting tests comparing blind to conventional reviews; (3) employ a scoring system in which proposals are rated separately on a number of criteria as well as overall; and (4) where feasible, formalize, or at least inform applicants of the importance of, any unwritten decision rules used by reviewers.

The Director of NIH should also (1) make greater use of subpanels and more fully integrate the work of mail reviewers into the panel process and (2) improve evaluation and oversight by retaining data on scores given by individual panelists and the race and gender of applicants. The Director of

NSF should also (1) increase the use of panels where feasible, (2) more closely monitor the inclusion of women and minorities among external reviewers, and (3) increase efforts to calibrate ratings among reviewers. The Chairman of NEH should also (1) improve the level of relevant expertise in reviews by making greater use of mail reviewers, (2) improve evaluation and oversight by collecting data on the race of applicants, and (3) increase efforts to calibrate proposal ratings among reviewers.

Agency Comments and GAO's Response

Officials of the Public Health Service (PHS), NSF, and NEH reviewed a draft of this report. All three agencies generally concurred with the need for increasing representation of younger scholars among reviewers and the need to identify any unwritten criteria reviewers may be using. PHS questioned whether the reliance of reviewers on preliminary results exemplified an unwritten rule at NIH. But NIH instructions to applicants describe the presentation of such findings as "optional," whereas GAO observed a universal expectation among reviewers that they be included in the proposal. PHS also agreed to consider using blind reviews to test for discrimination. Although NSF and NEH officials disagreed with this recommendation, it seems that they misinterpreted its contents; the recommendation has been clarified. All three agencies disagreed with scoring by dimension; however, they misinterpreted the recommendation as meaning weighted numerical scoring by criterion. This is not the case. What GAO is recommending is a procedure that ensures that all elements of a proposal are at least considered by reviewers in arriving at an overall appraisal.

NSF and NEH both agreed to improve calibration of ratings among reviewers. NEH argued that it already makes such efforts, but GAO did not observe them when attending NEH panels. NSF also agreed to monitor for race and sex discrimination among reviewers, although the agency cited legal prohibitions on collecting and retaining relevant data. GAO is aware of no such prohibitions, although it is true that reviewers may decline to provide such information. The fact is, however, that the data NSF supplied to GAO included information on the gender of reviewers. Finally, NEH disagreed with the recommendation to collect data on the race of applicants on the grounds that GAO had not found evidence of discrimination at NEH. This disagreement, however, is specious: the basis of the recommendation is not that GAO found discrimination but, rather, that there were no data to evaluate whether there was discrimination or not. Unless the data are made available, neither the agency nor the Congress can know whether such discrimination is occurring.

Contents

Tables

Figures

Abbreviations

GAO	U.S. General Accounting Office
NAS	National Academy of Sciences
NEH	National Endowment for the Humanities
NIH	National Institutes of Health
NSF	National Science Foundation
PHS	Public Health Service

Introduction

The economic and cultural leadership of the United States depends largely on the quality of its research in the sciences, arts, and humanities. Federal support for this research is often guided by a process of peer review to advise ultimate decisionmakers in the funding agencies. The essence of peer review is that the merit of a highly specialized proposal is evaluated by persons with appropriate expertise. Although scholars generally prefer peer review to alternative methods of research funding, they disagree about how well it works, how it should work, and what to do about its shortcomings. This controversy raises questions for the Congress regarding how well peer review assists in the allocation and management of federal funds.

Although researchers express some concerns about the efficiency and efficacy of peer review, the major concern among scholars, the one that generates the most letters to scholarly journals and to the Congress, is whether peer review is fair. A fair review is both expert and unbiased. The underlying integrity of peer review depends not only on the fact of its fairness but also on participants' perceptions that it is fair. One often perceived fundamental tension is that the highly expert reviewer may also be highly self-interested; that is, the closer a reviewer's work is to what is being reviewed, the better the reviewer may be able to judge its merits and the more likely he or she may be to care unduly about its success in winning funding. In the extreme, direct competition between an applicant and a reviewer could lead to the biased evaluation of a proposal.[1] However this tension between expertise and potential bias may be controlled or mediated by professional norms.

A second perceived problem is that reviewers are generally asked to judge not just the research design but also additional factors, such as whether the applicant's credentials and track record indicate that he or she is likely to successfully complete the proposed project or whether he or she has access to adequate facilities and institutional support to successfully execute the study. This broad addition of evidence beyond the research design has traditionally been seen as a prudent inclusion of all the evidence regarding the likely success of a proposal but could be construed as imposing an unfair disadvantage on applicants not already in the funding system (such as new scholars), those at less well known institutions, and those with fewer resources enabling them to interact with colleagues.

[1]This is why some critics argue that professional norms are more likely to work if the reviews are public, not confidential.

At the request of the Senate Committee on Governmental Affairs, we conducted a study of peer review in federal funding agencies. As agreed with the Committee, we focused on the peer review systems at three agencies: the National Institutes of Health (NIH), the National Science Foundation (NSF), and the National Endowment for the Humanities (NEH).

Background

The first recorded use of "peer review," or the evaluation of scholarly work by other scholars with the necessary expertise to judge its merits, is thought to have occurred in 1665, when the British Royal Society directed that the "Philosophical Transactions . . . be licensed under the charter of the Council of the Society, being first reviewed by some members of the same."[2] Since then, peer review has spread around the world and is widely used to advise on and legitimate funding and publication decisions in the sciences, arts, and humanities. However, there is no single agreed-upon formula for state-of-the-art peer review.

Today, agencies across the federal government use peer review to guide decisions awarding grants to individuals, universities, cultural institutions, corporations, and other entities. Agencies funding basic science and humanities research such as NIH, NSF, and NEH rely heavily on peer reviews, but each has its own distinct way of conducting them.[3]

Controversy Surrounding Peer Review

Peer review of proposals for federal research support has probably always aroused some criticism, reflecting the tensions within science and between science and society. With more scientists pursuing research support and the federal budget under increasing constraints, debate has reached new intensity in the last two decades. In general, complaints about peer review systems focus on three major issues: efficiency, in terms of the time and effort expended to seek funding; efficacy, in terms of providing the nation with the "best" science, humanities, and arts for its

[2]Harriet Zuckerman and Robert K. Merton, "Institutionalized Patterns of Evaluation in Science," in Robert K. Merton, The Sociology of Science (Chicago: University of Chicago Press, 1973), p. 463.

[3]In 1986, NSF began using the term "merit review" when referring to the evaluation of proposals submitted for funding. Former NSF director Eric Bloch argued that the term more accurately depicts the process of proposal evaluation at NSF, in which the proposal is reviewed first for its technical merits and then for additional characteristics such as its contribution to the research infrastructure and the goal of equity in the geographical distribution of funding. (See NSF Advisory Committee on Merit Review, Final Report (Washington D.C.: National Science Foundation, 1986).)

investment of tax dollars; and, equity, in terms of the fairness of the process for applicants.[4]

The principal criticism regarding efficiency is that gathering the volume of evidence and paperwork necessary to complete a proposal takes too much time, both for applicants to prepare and reviewers to digest, and that this diverts scientists from investigation to grantsmanship.[5] Scientists face strong incentives to master the fine points of proposal writing and lobbying at the expense of time and effort spent mastering their craft. The burdens on reviewers are also cause for concern. At NIH, for instance, an internal study found that reading proposals, traveling to Bethesda, and attending panels took from 30 to 40 days a year for participating panelists.[6] This amounts to significant time, with little direct compensation.

Concerns about efficacy refer to whether peer review leads to the best possible science, humanities, or arts. Many distinguished scholars have criticized peer review largely because of the tendency they see for it to minimize the very risks that new ideas almost always imply. Richard Muller and Michael Scriven have argued that peer review as currently practiced is highly risk averse, discourages interdisciplinary work, and penalizes scholars who explore new fields.[7] The Nobel laureate J. D. Watson recently emphasized the inherent need for researchers to take risks and explore new fields: "To make a huge success a scientist has to be prepared to get into deep trouble. . . . If you are going to make a big jump in science, you will very likely be unqualified to succeed by definition."[8]

[4]See Daryl E. Chubin and Edward J. Hackett, Peerless Science: Peer Review and U.S. Science Policy (Albany, N.Y.: SUNY Press, 1990), pp. 2-6. Peer review is also widely used by scholarly journals to evaluate articles on completed work and for many medical doctors and engineers and others to evaluate professional work and practices. The term peer review is even sometimes used to describe review by "in-house" experts. Although criticisms of one form of peer review may be germane to another, we focus primarily on critiques of peer review of grant proposals for federal support.

[5]Rosalyn S. Yalow, "Is Subterfuge Consistent With Good Science?" Bulletin of Science Technology and Society, 2 (1982), 401-4; Rustum Roy, "Funding Science: The Real Defects of Peer Review and the Alternative to It," Science, Technology, and Human Values, 10 (1985), 73-78.

[6]NIH, Sustaining the Quality of Peer Review: A Report of the Ad Hoc Panel (Bethesda, Md.: 1989).

[7]Richard Muller, "Innovation and Science Funding," Science, 209 (1980), 881; Michael Scriven, Evaluation Thesaurus, 4th rev. ed. (Newbury Park, Calif.: Sage, 1991). See section on shared bias.

[8]J. D. Watson, "Succeeding in Science: Some Rules of Thumb," Science, 261 (1993), 1812.

Another Nobel laureate, Rosalyn Yalow, has noted that peer review's need for public accountability can conflict with the freedom necessary to cultivate scientific breakthroughs.[9]

Although all three issues—efficiency, efficacy, and equity—are important as well as interrelated, the Committee asked us to focus primarily on the equity or fairness of the peer review process.

Perceptions of Bias

Many critics believe that in peer review a select group of scholars from a small number of elite universities repeatedly decide to fund one another's research while waving the flag of merit review to justify their funding decisions.

Charges that peer review is unfair have led to several surveys of participant perceptions. An independent survey of a small sample of applicants to NIH's National Cancer Institute found that 61 percent thought reviewers were reluctant to support unorthodox or high-risk research, approximately 40 percent agreed that initial review groups were controlled by an "old boys'" network, 34 percent agreed that reviewers were biased against researchers in nonmajor universities or institutions in certain regions of the United States, and 17 percent believed that reviewers are biased against young investigators.[10]

In a 1988 NSF survey of more than 14,000 investigators submitting proposals during the 1985 fiscal year, 38 percent indicated dissatisfaction with the peer review process. The most frequently volunteered reasons for their dissatisfaction, in order of importance, included (1) reviewers or panelists were not experts in the field (18 percent); (2) reviews were perfunctory, cursory, and nonsubstantive (17 percent) or conflicting (12 percent); (3) the peer review system was compromised by cronyism, politics, or an "old boys'" network (12 percent); and (4) the funding decision was unclear or inconsistent with the reviews (10 percent).[11]

[9]Yalow. She also made this point in her presidential address to the Endocrine Society; see Endocrine, 106:1 (1980), 413.

[10]Chubin and Hackett, p. 66.

[11]NSF Program Evaluation Staff, Proposal Review at NSF: Perceptions of Principal Investigators (Washington, D.C.: February 1988), p. 16.

Beyond Perceptions

Ironically, this debate has taken place in a near vacuum of empirical data available to independent evaluators.[12] Indeed, the lack of data on how official and unofficial criteria are applied or weighed has probably exacerbated the controversy by constraining it to intense perceptions and scant facts.[13] Moreover, outcomes are rarely studied, so little has been done to check the effectiveness of the agency grant criteria. That is, the criteria have not been shown to be empirically related to the production of good work.[14] Surveys of perceptions of the process have not helped much because they measure only perceived bias and cannot show actual bias or even influence; only by studying actual decisions can these issues be addressed.

A few studies do go beyond perceived biases to more rigorous evaluation. One of these, of peer review at NSF, found little evidence of actual bias. This study evaluated actual peer reviews rather than perceptions of the review process and analyzed decision factors by comparing blind and nonblind reviews of the same proposal. The findings suggest that reviewers from top-ranked university departments showed no favoritism toward applications from other top-ranked departments. Nor did reviewers seem unduly favorable to proposals from their own region of the country or biased against young investigators. Rather, the most important determinant of funding was the ranking of the proposal. However, the authors found a surprising degree of randomness in the scores. Indeed,

[12]Chubin and Hackett recount how difficult it is to enter the "black box" of peer review because the process and the data are shielded from the public eye (p. 80). They argue that there are few truly independent studies because "the process is at nearly all points inaccessible, opaque, and heavily infused with the values and interests of stakeholders" (p. 50).

[13]Without data, agency officials cannot know for certain what kind of problems various types of applicants are having. If, for instance, they discover that women are not scoring as well as men, they would not be able to tell if it is because women are submitting weaker research designs or if they are picking stale or overly novel research topics or if they are seen as getting less institutional support than men.

[14]Concerns about improving the efficacy of U.S. science and humanities funding are similarly constrained by a lack of data on the application of ratings criteria. That is, a given agency or program may be found to be more or less effective than others, but few data allow policymakers to learn why.

25 percent of the award-or-decline decisions were reversed by the blind second panel.[15]

Some observers believe that the in-house evaluations NSF and NIH conducted are little more than exercises in legitimation.[16] Nevertheless, some of the findings from these studies are hardly self-serving. For instance, the NSF study on perceptions of principal investigators found that, overall, 75 percent of all applicants had served as reviewers or panelists in the last 5 years; however, among consistently successful applicants, 97 percent had been reviewers.[17]

In spite of the in-house studies concluding that the research grant programs of NSF and NIH are basically fair, the debate regarding peer review has continued with an almost continuous series of critiques and rebuttals in scholarly journals such as Science, Nature, and The Journal of the American Medical Association. The Congress joined the fray again in 1979 with hearings by the House Committee on Science and Technology concerning allegations of fraud among scientists. The discovery of fraud previously undetected by peer review raised questions about peer review's credibility for authoritative judgment of the quality and accuracy of proposals. In 1980, the Congress asked us to evaluate the systems used to review proposals at NSF and NIH. Although we did make some recommendations that could increase accountability in proposal evaluation at NSF and NIH, we concluded that peer review in those agencies was free of widespread fraud and abuse.[18]

As for NEH, the controversy regarding peer review there has largely centered on charges of politicization, alleged to affect every stage of the process including reviewer selection, conduct of peer review panels, and final agency funding decisions. Some of the specific allegations concerned

[15]Stephen Cole and Jonathan R. Cole, Peer Review in the NSF: Phase Two (Washington, D.C.: National Academy of Sciences, 1981). Another analyst argues that such variance in the rating of proposals properly reflects differences in opinion, rather than random error, and that more homogeneous ratings would simply reflect a higher degree of shared prejudices. Stevan Harnad, "Rational Disagreement in Peer Review," Science, Technology, and Human Values, 10 (1985), 55-62. A similar debate recently took place in the British journal Nature. Ernst and colleagues published their findings regarding the irreproducibility of peer reviews of articles for journal application, and several readers wrote in to argue that the level of variance was reasonable and reflected a healthy level of debate. E. Ernst, T. Saradeth, and K. L. Resch, "Drawbacks of Peer Review," Nature, 363 (1993), 296. Letter responses and rebuttals are in Nature, 364 (1993), 183-84.

[16]Chubin and Hackett, p. 52.

[17]NSF, Program Evaluation Staff, p. 2.

[18]U.S. General Accounting Office, Better Accountability Procedures Needed in NSF and NIH Research Grant Systems, GAO/PAD-81-29 (Washington, D.C.: 1981).

intolerance of cross-cultural approaches, double standards being applied to third world perspectives, and blackballing of work on Latin America, some women's studies, and social change.[19] An important issue underlying all these complaints was whether or not "balance" was necessary within each proposal or merely in the overall portfolio of proposals.[20] The controversy led some scholars to conclude that the peer review process was being either trampled or ignored.[21] Other observers have countered that

"People who know the workings of . . . NEH say that the single most formidable obstacle to responsible allocation of grants is peer review. Rather than present the public's interests, too many of these panels represent their own artistic and scholarly cliques; they dole out money to allies and proteges, feather their own nests and keep it all in the family."[22]

In 1987, we offered descriptive data on the geographical distribution of this funding.[23] We reported that (1) the percentage of federal research funds received by the top-funded 100 universities had remained stable, although there had been considerable movement of universities in and out of the top 100; (2) federal research funding to universities and colleges was concentrated in relatively few states and institutions; and (3) state rankings on federal research funds to universities and colleges was correlated with size of population, number of employed scientists and engineers, number of Ph.D.s granted in science and engineering, state funding of higher education, and total federal research and development funds.

Panel Versus Ad Hoc Mail Review

Peer review by panelists may have several advantages over that by external ad hoc mail reviewers.[24] Panel reviewers may be less likely to be

[19]See, for example, Karen Winkler, "Humanities Agency Caught in Controversy over Columbus Grants," Chronicle of Higher Education, March 13, 1991, pp. A5-9; William McGurn, "Borking the Humanities," National Review, June 10, 1991, pp. 16-17; Stephen Burd, "Chairman of Humanities Fund Has Politicized Grants Process, Critics Charge," Chronicle of Higher Education, April 22, 1992, pp. A1 and A32-33; Stephen Burd, "Role of NEA, NEH Peer-Review Panels Questioned," Chronicle of Higher Education, May 20, 1992, p. A21.

[20]Karen Winkler. See also Lynne Cheney, "The Africans," Washington Post, October 14, 1986, p. A15.

[21]Stephen Burd, "Rift Grows Between Scholars and U.S. Officials Over Way Federal Funds Are Awarded," Chronicle of Higher Education, July 19, 1992, pp. A18-19.

[22]Jonathan Yardley, "Helms and the Art of Pragmatism," The Washington Post, July 31, 1989, p. C2.

[23]See U.S. General Accounting Office, University Funding: Patterns of Distribution of Federal Research Funds to Universities, GAO/RCED-87-67BR (Washington, D.C.: 1987). In another report, University Funding: Information on the Role of Peer Review at NSF and NIH, GAO/RCED-87-87FS (Washington D.C.: 1987), we provided descriptive information on NSF and NIH peer review systems used to fund university research and discussed equity issues.

[24]"External," "ad hoc," and "mail" reviewer are interchangeable terms.

working in precisely the same subfield as the applicant and consequently less likely to be direct competitors. Compared to mail reviews, panel reviews occur in a relatively public fashion in front of one's peers, and this provides at least a potential check on reviewer bias. Coming together on a panel allows reviewers to better calibrate their ratings and experience with many reviews, instead of just one or two through the mail; it also allows scholars to develop greater expertise in the art of reviewing.

Mail reviews have fewer, but nevertheless significant, potential advantages over panel reviews. Reviewer selection can be tailored to fit the field of the reviewer to that of the application, and the use of mail reviewers does not require the expense of travel and per diem costs.

Objectives, Scope, and Methodology

The congressional request and discussions with Committee staff led us to three objectives: (1) developing a framework to evaluate potential peer review weaknesses, focusing in particular on criticisms of fairness made by persons experienced in the process; (2) identifying current agency policies designed to address these potential weaknesses; and (3) assessing the extent to which some of the identified potential weaknesses in fairness actually occurred in peer review systems. With the agreement of Committee staff, we focused our work on the National Institutes of Health, the National Science Foundation, and the National Endowment for the Humanities.

We communicated the results of our work on the first two objectives to Committee staff in a briefing. We addressed the first objective by reviewing the extensive literature on peer review, interviewing participants, and consulting with experts. The framework we developed is shown in table 1.1. For the second objective, we reviewed agency manuals, internal studies, program descriptions, applicant and reviewer guidance, and application packets and interviewed agency officials. The results of that work are in appendix I.

Table 1.1: Potential Weaknesses in the Fairness of Peer Review

Decision stage	Potential weakness
Reviewer selection	Professional conflicts Lack of relevant expertise Personal familiarity with applicants Demographic and regional biases Halo effect
Reviews	Institutional and financial conflicts Unwritten criteria Inexpert reviews Matthew effect Personal familiarity with applicants Demographic and regional biases Halo effect
Funding	Excessive avoidance of risk Well-known applicants Matthew effect Demographic and regional biases Halo effect

In the rest of this report, we address the third objective: to assess the extent to which some potential weaknesses identified in addressing the first objective are actual weaknesses. We decided to examine potential problems of fairness in the peer review process and, based on these findings, to make recommendations on how those problems might be reduced. We looked at a wide array of data relating to specific proposals at the three agencies.

Peer review entails three discrete decisions: the selection of reviewers, the ratings made by reviewers, and the final awards by the agencies. Because there are allegations of unfairness or bias regarding all three decisions, we developed three evaluation questions for this objective: (1) To what extent do identified potential weaknesses in reviewer selection actually occur? That is, do reviewers possess relevant expertise in the subject of the proposal, and do they represent their applicant peers in terms of such factors as gender, age or seniority, region, and institutional affiliation? (2) What factors are related to how reviewers score proposals? (3) What factors are weighed in agency decisions to award funding?

Sampling at NIH, NSF, and NEH

Each agency considers thousands of applications for funding each year. To assess how well the peer review processes at these agencies are working, we first selected samples of new research applications at each agency, approximately equal numbers of which had been awarded or declined

funding.[25] This allowed us to make direct comparisons between the factors associated with successful and unsuccessful applications.

Because of differences in the program structures, data bases, and processes, we used somewhat different sampling strategies. At NIH, we first randomly selected five institutes: The National Institute of Allergy and Infectious Diseases; National Cancer Institute; National Heart, Lung, and Blood Institute; National Eye Institute; and National Institute of Diabetes and Digestive and Kidney Diseases. Then we randomly selected an equal number of new, regular research grant proposals, known as R01 applications, that were granted and denied funding in the fiscal year 1991 cycle.

At NSF, we first identified fields within each of the agency's major divisions that would allow us to combine NSF data with information on the relative rankings of applicants' and reviewers' employing institutions. Within those fields, we identified five programs from which to draw our sample of applications: biochemistry; economics; mathematics, algebra, and number theory; electrical engineering (solid state and microstructures); and theoretical physics.[26] Among these, we randomly selected a total of 50 successful and 50 unsuccessful research proposals from the fiscal year 1991 funding cycle.

We sampled research and fellowship applications from the three NEH divisions that most actively use peer review: Research Programs, Education Programs (Higher Education in the Humanities), and Fellowships and Seminars.[27] The processes in use at NEH resulted in our modifying the sampling of reviewers at that agency. There, each panel member reviewed all the proposals. In order to ensure that we did not unduly burden panelists, we determined that each of the 50 members of the 10 panels in the sample would be asked to respond to survey questionnaires on no more than 2 proposals, for a total of 100 questionnaires on 20 proposals. We then selected 34 additional proposals and surveyed the external reviewers.

[25]We occasionally refer to a "declination," or the failure of a proposal to obtain funding for any reason. Such a proposal may have been refused funding by the agency, deferred for future consideration, or withdrawn from competition. Often such proposals are resubmitted and funded in subsequent rounds of competition.

[26]In some cases, we selected programs in specific fields because they presented less demand for physical capital than most alternatives. Heavy capital requirements practically restrict the number of institutions that can plausibly compete for funding in some scientific fields; this would be the case, for example, with particle physics as opposed to theoretical physics.

[27]Because NEH programs place less emphasis on research grant support, a representative sampling of NEH programs had to include peer reviewed education and fellowship grants.

In some cases, analyses are based on larger sample sizes because data bases could not be disaggregated. This is indicated by the sample sizes shown in the relevant tables and figures.

Data Sources

We used multiple sources of data relating to the three agencies. First, we obtained extensive data from administrative files on selected programs at each agency. This information included background data about the applicants (or principal investigators), panelists and external reviewers, the requested funding amounts, the scores given by reviewers, and the agency's ultimate decision on whether to fund the projects and, if so, the amounts granted. Because computerized data bases were less complete at NEH than at the other agencies, we had to collect large amounts of data from paper records. Therefore, we conducted our field work first at NEH, completing our manual data collection before fiscal year 1991 data were available. The data for this agency are from fiscal years 1989 and 1990.

Next, we supplemented these data by surveying panelists and external reviewers for each of the proposals in our agency samples. Our survey was administered to a total of 1,370 reviewers from late May to September 1992, with 1,181 responses, for a response rate of 86 percent.

Each questionnaire had four sections. First, we provided the reviewers with the applicant's name, department, institution, and proposal title, and then we asked reviewers about their knowledge of the topics involved in this proposal and their perceptions of the applicants' institutional affiliations and contributions to the field. In the second and third sections of the questionnaire, we asked the reviewers for some general information about themselves and how well they knew the applicant and program officers. In the fourth section, we invited reviewers in open-ended questions to comment on both the benefits and drawbacks of the agency's peer review practices and what changes they would recommend.

In addition, for NIH we used the survey to obtain the scores of individual panelists on the sample proposals. Using the NIH data base, we could not identify which panelists had been primary or secondary reviewers of individual proposals. Therefore, we deliberately oversampled these panelists to maximize the probability of sampling at least one such reviewer for each proposal (for a total sample of 695). In fact, we succeeded in obtaining scores for 99 of the 100 cases.

The NSF data base did include individual scores, so we did not ask respondents from that agency questions on this item. However, some of the panelists included in our survey had not been primary or secondary reviewers, so they could provide only limited information. Nevertheless, we had usable reviewer responses for 95 of the 100 sampled applications. At NEH, our total sample was 277 and yielded responses on 52 of the 54 proposals.

In addition to agency and survey data, we obtained measures of the prestige of the applicants' and reviewers' home institutions, where available. Our major source for most academic departments was the National Academy of Sciences' (NAS's) 1982 rating of faculty quality.[28] For departments in professional schools of medicine, dentistry, and veterinary medicine, we relied on the National Education Standards Gourman Report.[29] Although the NAS study was 10 years old when we did our work, it is generally regarded as the best available source. The Gourman Report, published in 1989, is more current.

We also observed a number of peer review panels at each of the three agencies during the consideration of fiscal year 1993 grant applications. This allowed us a fuller understanding of the dynamics of panel behavior, including the variations in procedures, tone, and substance at different panel meetings. In addition, we observed a meeting of one of the NIH institute councils, including both public and closed sessions.

Ideally, we would have preferred to be anonymous at these meetings, but this was not generally possible. In most cases, panelists were informed that we were present. Even where this was not the case, however, we cannot be sure that panelists were unaware of our presence. Several science-related publications gave extensive coverage to our ongoing work, raising fears about possible deleterious effects on peer review that could ensue. One researcher wrote a critical letter to the Committee on Governmental Affairs, which was the basis for an article in Science.[30] A letter to GAO criticizing the study was discussed in another publication distributed to NIH panelists in their meeting packets.[31] And many panelists undoubtedly saw an article that portrayed the study as potentially

[28]Lyle V. Jones, An Assessment of Research-Doctorate Programs in the United States, vols. 1-5 (Washington D.C.: National Academy of Sciences, 1982).

[29]Jack Gourman, The Gourman Report (Los Angeles, Calif.: National Education Standards, 1989).

[30]Elliot Marshall, "An NSF Survey Rattles Some Nerves," Science, 257 (1992), 620-21.

[31]"GAO 'Survey of Research Proposal Reviewers,'" NIH Peer Review Notes, October 1992, p. 2.

threatening to the research community and that was accompanied by the illustration shown here (see figure 1.1).[32] Knowledge of our presence may have affected the behavior of participants, so that our observation of any problems at panel meetings may be regarded as conservative.

[32]Bruce Agnew, "Looking for Mr. Goodpeer: GAO Eyes Peer Review," The Journal of NIH Research, 4 (October 1992), 42-43.

Figure 1.1: Looking for Mr. Goodpeer

Source: Drawing by Andy Meyer, reprinted from Journal of NIH Research, 4 (October 1992), p. 42.

Strengths and Limitations

This study has a number of major strengths. First, because of our unique access to data from most federal agencies, we were able to go beyond participants' perceptions of peer review to look at specific funding decisions across agencies. In contrast, most cross-agency studies have relied almost exclusively on perceptions of program officers, other agency officials, reviewers, and applicants, while studies using actual data on individual cases have been typically in-house at individual agencies. Thus, our strengths lie in having done an external evaluation of in-house data and in having concentrated on what actually happened rather than on perceptions of what happened.

Second, by bringing together multiple data sources, including agency data, survey information, and direct observations, we were able to examine many of the influences thought to affect the peer review process and to test whether those influences could be demonstrated empirically. Using agency data alone would not have permitted us to conduct these analyses.

Third, the comparative framework we adopted helped us identify both the common traits of different peer review systems and the specific circumstances of individual agencies. This approach allowed us to develop recommendations properly pitched to the general or particular levels, as appropriate.

Our study has three principal limitations. First, because of the differences in organization, processes, and data bases, we could not use the same sampling techniques in all three agencies, nor are all the data precisely comparable. This limits our ability to generalize our results even across the three agencies. Nevertheless, the data we do have are quite similar in most cases, and where they are not, we note that fact in the body of the report.

Second, as with any study of social behavior, we are constrained by limitations on what we can measure, so we present analyses that show the empirical relationships between variables describing applicants and the scores reviewers give to their proposals, but we cannot know for certain how the reviewers weighed those factors nor what other, unmeasured, factors may have affected their decisions. For instance, in all the programs we studied, only a single summary score is given. Therefore, and this is a crucial point, neither we nor the agencies know how the reviewers rated the quality of the research design or the importance of the research question. In addition, we have only limited information on the personal and professional relationships among applicants, reviewers, and program

officers; although such relationships may be important, gathering solid empirical information on a large number of cases is extremely difficult.

Third, we can shed some light on what factors are related to peer review decisions, but in some cases these factors are quite controversial. Normative or fairness issues about whether to apply a given criterion cannot be readily resolved by empirical analysis. One person's prudence may be another's prejudice. For instance, reviewers or program officers might be more ready to fund a large project for an applicant whom they know than one they do not. Or they might be more inclined to fund someone with a great publication track record. Participants can construe each of these criteria either as prudent financial management and risk-averse behavior or as bias against less well known scholars.

We had an additional problem at NIH, where the data on the scores given by individual reviewers to specific proposals were not available and had to be collected with our survey. Since these data are based on recall, they may be less reliable than the NSF and NEH administrative records. However, many of the NIH reviewers who called us with questions about the survey reported having detailed records of all their scoring and review decisions. For these respondents, at least, recall was not an issue. The questionnaire also instructed reviewers who could not recall their reviewing of the proposal to skip all questions specifically relating to the proposal.

We conducted our field work in accordance with generally accepted government auditing standards between March 1991 and October 1993. Statistical relationships discussed in the text are significant at the .05 level unless otherwise noted. We obtained written comments on a draft of this report from the Public Health Service (for NIH), NSF, and NEH.

Organization of This Report

In the remainder of this report, we address the three evaluation questions listed under the third objective of our study. In chapter 2, we examine reviewer selection—that is, the extent to which the reviewer selection process yields individuals who both possess the relevant expertise to judge the proposals they consider and represent their applicant peers in terms of such factors as gender, age, region, and institutional affiliation. In chapter 3, we provide an analysis of the factors associated with the scores reviewers give to proposals. In chapter 4, we consider the factors that influence agency decisions to fund (or not fund) individual grant applications. Finally, in chapter 5, we present our overall conclusions and recommendations.

Appendix I is our framework of potential weaknesses and agency actions and policies, while appendixes II, III, and IV contain comments from PHS (for NIH), NSF, and NEH and our responses. Other appendixes list reviewers of this report and major contributors.

Selection of Peer Reviewers

The review of research proposals by peers is regarded as essential at the three agencies we studied, so the first major question we asked was, "To what extent do identified potential weaknesses in reviewer selection actually occur?"[1] Some critics of peer review recite horror stories of reviewers who know little or nothing about the problems addressed and the questions posed in applications or, conversely, of reviewers whose interests are so close to the applicants' that they can (and do) sandbag proposals from rivals. Others paint a picture of peer reviewers as predominantly senior-level white males from elite institutions, especially in the northeast. In short, many critics see peer review as rife with incompetence, conflict of interest, and favoritism.

In our analysis, we set out to evaluate the extent to which these perceptions are supported empirically. We combined the responses of each of the reviewers in our survey with agency administrative data and rankings of academic departments to examine (1) the extent to which those reviewers appeared to be knowledgeable about the subject matter in the proposals they reviewed, (2) the extent of personal and professional relationships between reviewers and applicants, and (3) how representative reviewers were when compared to applicants on such factors as gender, geographic area, academic rank, and prestige of their employing departments.

Policies and Procedures for Selecting Peer Reviewers

Each agency we examined uses its own peer review processes. In some cases, the agencies rely on panels of experts convened at a given location; in others, they use external reviewers selected in an ad hoc manner who provide reviews by mail. Although there are some similarities across the agencies, the differences are substantial and may affect ultimate decisions on who gets funding. Thus, in this chapter and chapters 3 and 4, we preface our analysis with a brief description of the relevant peer review policies at each agency. Here, we describe how peer reviewers are selected at NIH, NSF, and NEH.

There are two basic types of peer reviewers: panelists and external reviewers. Panelists perform their final reviews while meeting with colleagues on either ad hoc or sitting panels. External reviewers mail in

[1]Agencies generally track the gender, age or year of degree, type of degree, degree granting university, state, and present employer of reviewers. However, they do not have data pertinent to much of the controversy regarding peer review. For instance, they do not track how close the reviewer's area of expertise is to that of the subject of the proposed grant applications, nor do they track how well the reviewers know the applicants. They have not looked at differences between external and panel reviewers. (They also do little, with the partial exception of NSF, to match reviewer and applicant data.)

their reviews and do not attend a meeting. They are typically chosen ad hoc by the subject matter of the proposals submitted. We use "peer reviewer" to refer generically to all reviewers, whether panelists or external reviewers.

Reviewer Selection at NIH

At NIH, the Division of Research Grants has organized approximately 100 initial review groups headed by scientific review administrators, who are generally career employees with Ph.D.s and prior experience in research. Panelists are selected to cover a range of area expertise on any given panel. The initial review group panels review the proposals and send their recommendations to the institutes and their advisory councils.

The general criteria NIH uses to measure panelist expertise include Ph.D., M.D., or both; publications; honors; and seniority. Representativeness is addressed in reviewer selection by criteria that call for "adequate" representation of women and minorities, by rotating one fourth of the panelists each year, by conducting active outreach programs, and by publishing the names of panelists twice a year. Since NIH uses relatively few external reviewers, far fewer scholars serve as peer reviewers each year at NIH (7,400) than at NSF (60,000).[2]

Panels at NIH tend to be relatively large compared to those at the two other agencies. A typical panel at NIH includes 18 to 20 panelists; larger panels run up to 50 members.[3] At NSF, panels are typically constituted of 8 to 12 members, and at NEH they usually have 5.

Reviewer Selection at NSF

Three methods of review are used at NSF, including ad hoc external review, panel review, and a combination of mail and panel review. Each field has its own peer review tradition, and that, rather than any NSF-wide criteria, determines the method of peer review used in each program. Each of the three peer review formats covers about one third of NSF's programs.

In general, each academic science discipline has a corresponding NSF research grant program administered by a program officer who may be a career NSF employee or an outside scholar temporarily serving at NSF. Both

[2]NIH does use prior panelists with valuable subfield expertise to complement their standing panels through their reviewer reserve system. Occasionally, NIH does use an external review or, even more rarely, teleconferencing. Teleconference reviews have the advantage of allowing some cross-examination of the external reviewer.

[3]The larger panels usually break down into subpanels, but these subpanels can still be very large.

external reviewers and panelists are selected by the individual program officers from lists of potential reviewers that they develop and maintain.[4] Guidelines stress expertise, demonstrated ability, and general knowledge of the field. The race, gender, and age of reviewers is considered only after expertise is established. NSF guidelines also call for representation of small, medium, and large institutions, as well as nonacademic scientists. Persons who select panelists are expected to avoid concurrent or successive appointments from the same institution.[5] About 60,000 reviewers are used annually from a total of 150,000 reviewers, whose names are on lists kept by the various program officers.

Reviewer Selection at NEH

Peer review at NEH generally employs a single panel review system as opposed to the standing panels used at NIH and often at NSF. Applications are sent to the program manager of the appropriate subject area program unit. After receiving the bulk of the proposals, the program manager tries to customize the make-up of panels to match the current crop of proposals. These are small review panels, typically including five specialists from outside the agency. The evaluations by the panels are sometimes supplemented with individual reviews or independent letters of reference solicited by the program officer from external specialists in the subject area. NEH uses peer review for all grant programs. All research grants are funded by the division of research programs, which also uses prepanel specialist external reviewers.[6]

About 1,000 scholars a year, drawn from a computerized list of 13,000 names, serve on approximately 150 panels. Expertise is established by a reviewer's track record in publications, exhibits, and films. Representativeness guidelines call for panels to reflect cultural and geographic diversity, and there are caps on how many times any individual can sit on a panel.

Common Policies, Problems, and Recent Initiatives

All three agencies have policies to prevent financial and institutional conflicts of interest. The potential for institutional bias on the part of reviewers is controlled at the selection stage and during panel meetings, when reviewers from the same institution as the applicant are asked to leave the room while that proposal is discussed. (The implementation of

[4]No one system of panel member rotation is used at NSF.

[5]All agencies have had to define "institution" carefully because some large universities have multiple campuses. Because of representativeness and conflict of interest issues, each agency has its own explicit definition of what constitutes the same institution.

[6]NEH's division of preservation and access also makes extensive use of prepanel reviewers.

these policies of temporary recusal from panels is addressed in chapter 3.) Financial conflicts of interest are universally addressed by having all reviewers sign a statement denying financial conflict.

However, "intellectual camp" conflicts are less rigorously screened. By camp conflicts we refer to cognitive conflicts among scholars in a field. Typically these are conflicting views from the different sides of an academic debate.[7] Camp conflicts may be harder to measure and screen than financial or institutional conflicts but pose a potential conflict of interest because careers and reputations can be based on their outcomes. The agencies depend on their program officers and other panelists to screen these conflicts through their knowledge of the discipline and its debates; however, in our panel visits, we witnessed the fact that several camp comments went unchallenged.[8] The agencies also rely very heavily on self-report by the reviewers themselves with respect to their own conflicts and intellectual prejudices and passions.

These are weak controls compared to those for financial and institutional conflicts. For instance, the back of the NSF proposal evaluation form has a conflict-of-interest statement but it explicitly tells the reviewers that "regardless of any such affiliations or interests, unless you cannot be objective we would like to have your review." According to the former director of NEH's division of research programs, NEH's conflict-of-interest policy "does not consider personal animosities or conflicts based on differences of professional opinion."[9] However, conflicting ideas are often at the heart of scholarly inquiry, and their rigorous elimination might unnecessarily disqualify good reviewers who would give fair reviews in spite of their intellectual passions.[10]

NSF recently started allowing the principal investigator to "suggest" potential reviewers. These suggestions can be positive or negative, as

[7]These have also been referred to as invisible colleges. See Daryl E. Chubin and Edward Hackett, Peerless Science: Peer Review and U.S. Science Policy (Albany, N.Y.: SUNY Press, 1990), p. 80.

[8]The following comments by panelists went unchallenged by program officers or fellow panelists: "I have philosophical problems with that whole project"; "I do not like the post-structuralist approach"; "I am trying to overcome my prejudice against Midwestern literature"; "He's lefty trendy"; "I don't like this approach; I prefer a Markov"; "My three equals taste; this is a loser strategy and a bad line of research, although he is in the top four in this dubious field of research."

[9]Steven Burd, "Chairman of Humanities Fund Has Politicized Grants Process, Critics Charge," The Chronicle of Higher Education, April 22, 1992, p. A33.

[10]Shared biases are also a potential problem, so excessively rigorous screening of potential camp conflicts could inadvertently lead to creating the opposite problem by selecting only reviewers who share an applicant's biases.

when an applicant requests that a particular individual not be asked to review his or her proposal. The policy allowing such suggestions is intended to reduce the influence of camp conflicts and increase the relevant expertise of reviewers. Although allowing applicants to suggest possible reviewers could achieve these ends, it could also increase the risks of cronyism. NEH has a similar policy, but NIH does not. NEH allows for positive suggestions for external reviewers. Typically the program officer sends out five to seven invitations to review a proposal and tries to include at least two of those suggested by the applicant.

A common problem in the selection of peer reviewers is that persons selected and asked by the program officer to serve can and do decline. Therefore, underrepresentation does not necessarily imply program officer bias. Program officers we interviewed were concerned that the rate at which reviewers decline to serve on panels or return written reviews was rising. Moreover, these officers told us that they perceived a higher rate of declination among researchers at top universities than at lesser institutions.[11] The officials in charge of recruiting reviewers told us that the critics have it backward and that elite investigators and institutions are underrepresented rather than overrepresented in the peer review process.

Reviewers' Knowledge About and Interest in Proposed Work

Our first issue of fairness concerns the extent of reviewers' knowledge and interest in the research area, constituting a dilemma for peer review. On the one hand, peer reviewers should be informed on the subject of the proposal in order to judge it fairly. On the other hand, the more closely a reviewer's expertise matches an applicant's, the more likely it is that the two could be direct competitors or allies. As one NIH scientific review administrator told us, "We walk a fine line to get qualified reviewers without a conflict of interest." Although there are few documented cases of this, an opportunistic reviewer could sabotage a competitor's proposal or unfairly assist a friend. This tension is inherent in the peer review process. Consequently, program officers have incentives to recruit a variety of participants, some whose expertise is highly relevant and others who are not directly working in the area and can act as a check on bias toward a subfield.

[11]Similarly, Harold Varmus, the new director of NIH, in a recent interview said that one of his concerns about peer review at NIH was "the unwillingness of talented people to serve on study sections." "Varmus: The View From Bethesda," Science, 262 (1993), 1364.

Proximity of Reviewer's Expertise to Applicant's Proposed Work

Our survey of peer reviewers was inclusive because we drew our sample from all participants who reviewed or rated each proposal, not just primary reviewers.[12] In our questionnaire, we asked the respondents to comment on one or two proposals they had reviewed for the agencies. To determine the proximity of expertise, we asked them, "How close or distant is your own research to the research of the proposal you reviewed?" Table 2.1 summarizes the reviewers' responses for each agency. As we noted earlier, all relationships discussed in the text are significant at the .05 level, unless we indicate otherwise.

Table 2.1: Proximity of Reviewers' Expertise to Applicants' Proposed Work[a]

Agency	Same question	Related question	General area	Unrelated area
NIH	[b]	14%	52%	34%
NSF	5%	39	49	7
NEH	2	28	49	21

[a]Survey question: How close or distant is your own research to the research of the proposal you reviewed? Sample sizes are NIH = 215, NSF = 263, NEH = 227.

[b]Less than 0.5 percent.

NIH reviewers were much less likely than their counterparts at NSF and NEH to regard themselves as working on the same or a related question. In fact, more NIH reviewers, one third, said their work was unrelated to that in the proposal they reviewed than those from NSF and NEH. Several factors could contribute to these differences. First, as noted above, the programmatic divisions of NSF closely follow the disciplinary boundaries of academia. This is less the case at NIH, where basic science and applied medical questions overlap and have to be evaluated together and where participants come from both the colleges of arts and sciences and medical schools. At NEH, panels can represent a very broad spectrum of the humanities because it is a small agency that supports numerous disciplines, and the disciplines themselves cover worldwide topics and all periods of civilization. So the differences in the fields of research and the program lines of organization in the agencies may contribute to the observed differences in reported research proximity.

Second, NSF and NEH also make greater use of external reviewers, which allows for more precise matching of reviewers and proposals. Third, the size and scope of panels may affect the proximity of expertise. Panel size averages from about 20 at NIH to 5 at NEH. In all three agencies, very few

[12]We oversampled to ensure that we included a primary reviewer for most proposals.

reviewers reported working on the same question as the applicant. Because self-interest tends to increase with proximity, this virtual absence of highly proximate reviewers means that direct conflict of research interests was probably rare. However, the great majority of our respondents were working in a related field or the same general area as the applicant.

Reviewers' Knowledge of Relevant Literature

Our survey had another measure of substantive expertise. We asked reviewers, "Are you sufficiently familiar with the literature that you could suggest references that should be cited?"[13] Table 2.2 shows the same pattern as table 2.1, with NSF reviewers most able to cite references, followed by NEH and NIH. NIH had half as many reviewers who reported knowing the literature related to the proposal well. Furthermore, 79 percent of NIH, 57 percent of NEH, and 48 percent of NSF peer reviewers could cite no more than a few references. However, it can also be said that 72 percent of NIH, 85 percent of NEH, and 92 percent of NSF peer reviewers had at least some familiarity with the literature.

Table 2.2: Reviewers' Familiarity With Relevant Literature[a]

Agency	Yes, many such references	Yes, but only a few	No
NIH	21%	51%	28%
NSF	51	41	7
NEH	43	42	15

[a]Survey question: Are you sufficiently familiar with the literature that you could suggest references that should be cited? Sample sizes are NIH = 213, NSF = 261, NEH = 208.

Overall, the picture that emerges from tables 2.1 and 2.2 is mixed. Although most reviewers reported expertise in the general areas of the proposals they reviewed, many were not expert on closely related questions and could cite only a few, if any, references. This lack of proximate expertise was most pronounced at NIH. However, although this raises questions about the relative adequacy of NIH reviews and ratings, the greater proximity of NSF reviewers makes them potentially more vulnerable to apparent or actual self-interest in their reviews.

Our survey included several open-ended questions. For instance, we asked NIH reviewers, "In your view what are the benefits of the peer-review

[13]Some might argue that the question on knowledge of the literature does a better job of capturing current expertise and screening out those who have switched fields than self-ascribed research proximity.

process at NIH" and "In your view what are the drawbacks of the peer-review system at NIH?" Reviewers most frequently mentioned expertise as the most important (NSF and NEH), or second most important (NIH), benefit of the agency's peer review process, but lack of appropriate expertise was, aside from complaints about overwork and lack of money, the problem most frequently mentioned by NIH reviewers (28 percent). Lack of expertise was also one of the most frequently cited drawbacks at NEH (17 percent), but only 5 percent of NSF reviewers cited this as a problem. In sum, the reviewers' comments paralleled what we found in our survey questions.

Personal and Professional Relationships Between Reviewers and Applicants

Reviewers' Personal Knowledge of Applicants

Another commonplace allegation is that peer review is biased by a network of personal and professional relationships. Actual networking would have several dimensions that would include relationships between program officers and reviewers, applicants and reviewers, and reviewers and reviewers. These relationships would vary in closeness and longevity. We focused on the relationship most likely to influence scores: reviewer knowledge of applicants.[14]

Table 2.3 shows the extent to which reviewers directly knew the applicants whose proposals they reviewed or had indirect knowledge of the applicants through someone else. Our measure of direct knowledge was the following survey question: "Before your review of this proposal, were you and the applicant sufficiently acquainted that if you passed each other on the street you would be expected to stop and chat?" Our measure of indirect knowledge was an answer of no to the question above and yes to the following: "Say you had wanted to find out more about the applicant (for purposes unrelated to the proposal). Was there someone you knew and could talk to who had more knowledge about the applicant?" The

[14]We also asked all our respondents if they knew the program officer before they reviewed the proposal. We did not ask panelists if they knew the program officer before they were selected to be panelists.

latter question casts a very wide net, so it would be expected that many reviewers would have at least this sort of indirect familiarity with applicants.

Table 2.3: Reviewers Knew Applicants Directly or Knew Someone Who Did[a]

Agency	Direct	Indirect	Neither
NIH	23%	54%	24%
NSF	46	40	14
NEH	9	60	30

[a]Sample sizes are NIH = 213, NSF = 260, NEH = 228.

The majority of reviewers did in fact have either direct or indirect knowledge of applicants at all three agencies. However, differences are striking in the degree of direct familiarity. NSF reviewers were more than twice as likely as NIH reviewers, and more than five times as likely as NEH reviewers, to report direct personal knowledge of the applicants whose proposals they reviewed. Recall that NSF reviewers also had more proximate expertise than those at the other agencies. What we see here is the tension between proximate expertise and potential for bias, either for or against, because of personal familiarity.

Reviewers' Professional Relationships With Applicants

Personal and professional familiarity does not, of course, necessarily imply that the reviewer and investigator are either old friends or old enemies. Professional researchers devote their careers to mapping the frontiers of their fields, and debate regarding the validity of these efforts is central to the scholarly enterprise. Although such debates are about ideas, a researcher's career turns, in part, on how well peers are persuaded by his or her research and ideas. Thus, reviewers and applicants may have serious professional conflicts of interest that can be quite particular—as over an evaluation an applicant has given of the reviewer's work in the past—or quite broad, where theoretical, methodological, or philosophical issues divide a discipline into different intellectual camps. (Of course, professional conflicts can have personal overtones.) The extent of these conflicts is hard to determine because people may not give frank answers. As a result, the responses to our more pointed survey questions in this area may well underreport the degree of professional conflict.

In our survey we asked, "Is the focus of your research so similar that some scholars would believe you and the principal investigator are competitors?" The overwhelming majority said no, but reviewers for NSF

were more likely (10 percent of all respondents) to respond yes than those for NIH (1 percent) or NEH (3 percent).

When asked in our survey if the applicant had ever reviewed one of the respondent's grant proposals, few reviewers said yes, but NSF reviewers were more likely than others to respond affirmatively—8 percent versus 2 percent for NIH and 4 percent for NEH.[15] Similarly, when asked if the principal investigator had ever published a review of one of the respondent's publications, 7 percent of the reviewers for NSF said yes, compared to 2 percent for NIH and 5 percent for NEH.

Finally, in our section for open-ended comments, reviewers reported a broad range of drawbacks, with professional and intellectual camp conflicts among the most frequently reported concerns at NSF (16 percent of the respondents) and NEH (10 percent).

Overall, the vast majority of our respondents, even at NSF, said they did not have these potential professional conflicts, but where conflict of interest is concerned, even these small percentages reporting potential conflict are noteworthy. Reviewers at NSF more often reported that they could be seen as competitors, were more likely to report a past review of their work by the applicant, and were most concerned about camp conflicts. The close parallel between professional boundaries and agency program division boundaries at NSF may account for this trend. Currently, these potential conflicts are screened only ad hoc, with the applicant or reviewer left to raise the issue, or when a peer or the program officer knows about it and intervenes.

Representativeness of Peer Reviewers

Each of the agencies has policies to promote reviewer selection that is balanced in terms of race, gender, and region. NSF and NEH also have policies to promote age balance. So the third issue we address is that of representativeness among peer reviewers. The American Heritage Dictionary defines peer as "A person who has equal standing with another, as in rank, class or age." By "representativeness," we refer to the extent to which reviewers are peers of applicants. However, such strict equality is unlikely because peer reviewers are expected to be both peers and experts. Obviously, there can be tension between the criterion of equal standing and the superiority necessary for expertise. A more realistic expectation for peer reviewer selection is that there should be no

[15]Since the agencies do not release information about who reviewed specific proposals, this question probes reviewers' perceptions about who reviewed their proposals. Of course these perceptions can be wrong.

preference on the basis of race, ethnic identity, gender, region, age, or institutional affiliation. For example, expertise takes time to establish; reviewers should not therefore be expected to perfectly reflect applicants by age or academic rank.[16]

Reviewers might be expected to represent several possible peer groups, such as the set of all people with Ph.D.s or M.D.s in a given field or, more narrowly, all active researchers in the field or, still more narrowly, all applicants for funding under the same program; some would further limit the pool just to respected experts as the ideal. The broader definitions of "peer group" may be influenced by factors that have nothing to do with peer review. For instance, if we compared the region of reviewers to all Ph.D.s, we might also capture the unwanted influence of differential employment rates of Ph.D.s, across regions. So if more Ph.D.s in New England are driving cabs than Ph.D.s elsewhere, it is hardly appropriate to include them as peers. To minimize such extraneous factors, because program officers commonly add applicants to their lists of potential reviewers, and because the notion of reviewer representativeness—within an equity context—must inevitably be examined relative to applicants, we chose current applicants as the most relevant comparison group. (The most restrictive definition of peers as composed only of respected experts could exclude all sorts of newcomers and would be very hard to measure.)

We also compare reviewer and applicant distributions across agencies. Comparing applicants across agencies allows us to observe how the applicant base varies across fields, and comparing reviewers by agency can inform us about differences in agency preferences. Because of data limitations at the agencies, we could not examine reviewer selection by race or ethnicity, but we did review general representativeness by gender, region, age, rank, and institutional affiliation.

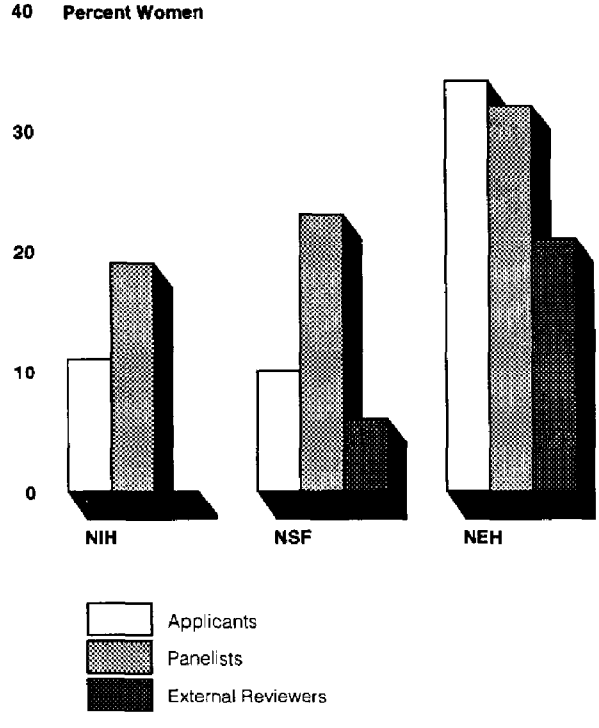
Gender Representativeness of Reviewers

In our sample, women were well represented on panels but not among external reviewers. Figure 2.1 shows that, compared to applicants, women were overrepresented on panels at NSF, and at NIH and NEH the percentage of women panelists approximated that of women applicants. Yet, at NEH, women apparently were underrepresented among external reviewers. (As already noted, NIH did not provide us with comparable data on external reviewers.) This is largely because all the external reviewers are used by the research programs division, which in fact had far fewer women

[16]Of course, some would contend that these representativeness criteria are functionally irrelevant and that a peer should simply be defined as someone with the knowledge to be a competent reviewer.

applicants than other NEH programs. So at NEH, the lack of women as external reviewers is less a factor of reviewer selection and more a reflection of the underlying pool of applicants.

Figure 2.1: Gender Representativeness at NIH, NSF, and NEH[a]



[a]There are no data on NIH external reviewers. Sample sizes are NIH applicants = 92, NIH panelists = 603, NSF applicants = 91, NSF panelists = 84, NSF external reviewers = 225, NEH applicants = 323, NEH panelists = 50, NEH external reviewers = 441.

However, at NSF a more complicated pattern emerges. The percentage of applicants who were women varied greatly by program. Some programs had very few women applicants, but, even so, women were underrepresented among external reviewers. Overall, in the five NSF programs we studied, external reviewers, but not panelists, were less likely to be women than were applicants. Such differences are hard to explain as the result of a differential acceptance rate. It may be that the more public and visible nature of panels serves to drive program officers toward a more balanced selection. Or it could be that in the process of assembling a panel, program officers simply are more conscious of the

gender distribution of the group than they are with ad hoc outside reviewers. In fact, when we asked NSF officials about this, they told us that they had been "putting pressure" on program officers to include more women on their panels but had not extended this to external reviewers.

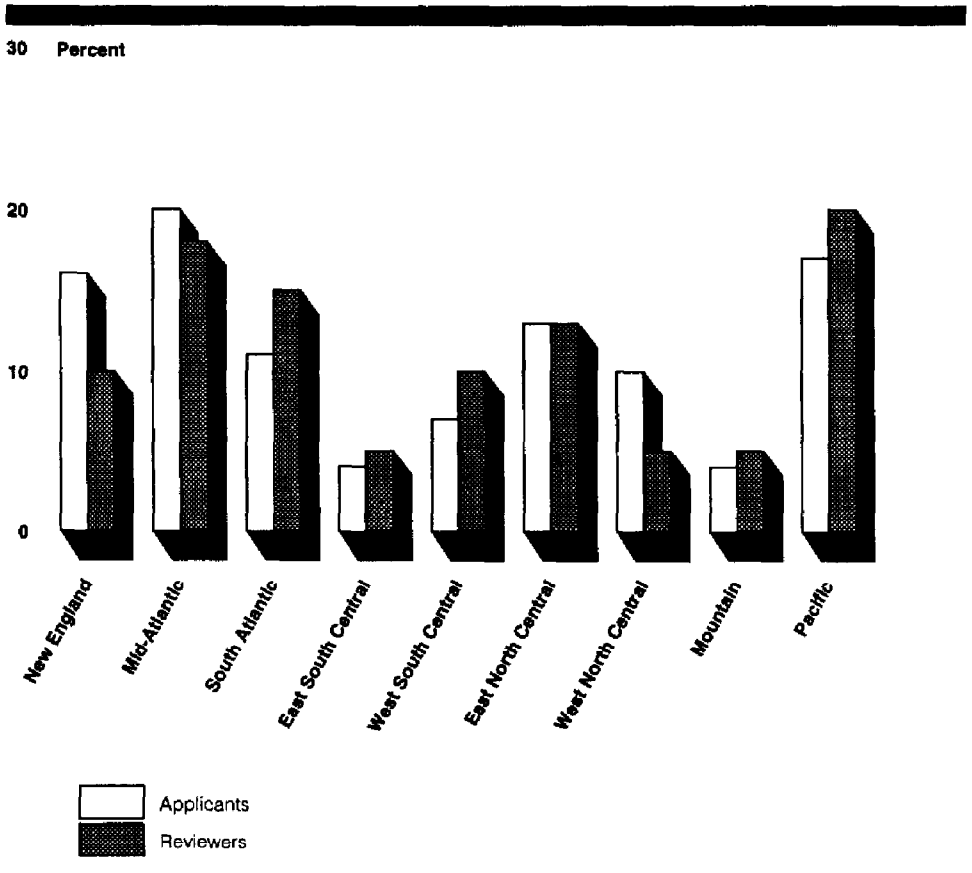
For several reasons, the underrepresentation of women among outside reviewers is a much more important problem at NSF than at NEH. First, only 6 percent of NSF reviewers were women, compared to 21 percent at NEH. Second, many NSF programs, such as mathematics, use only external reviewers. In contrast, NEH uses external reviewers only for prepanel review, so it has its representative panels to check and balance its external reviewers.

Regional Representativeness of Reviewers

One common complaint is that most reviewers represent only a few states and regions, particularly the Northeast. Others counter that universities are concentrated regionally. To examine regional representativeness, we compared the percentage of applicants to that of reviewers for nine regions.[17] Figures 2.2, 2.3, and 2.4 summarize these data for NIH, NSF, and NEH, respectively. Reviewers were representative of applicants at NIH, where only one region, West North Central, was underrepresented. (By underrepresented, we mean that a given region had a lower percentage of peer reviewers than applicants.) Reviewers at NSF and NEH were also representative of applicants' regions. At NSF, only the South Atlantic region was highly underrepresented, with about half the proportion of reviewers as applicants. At NEH, there was virtually no disparity in the regional distribution of applicants and reviewers.

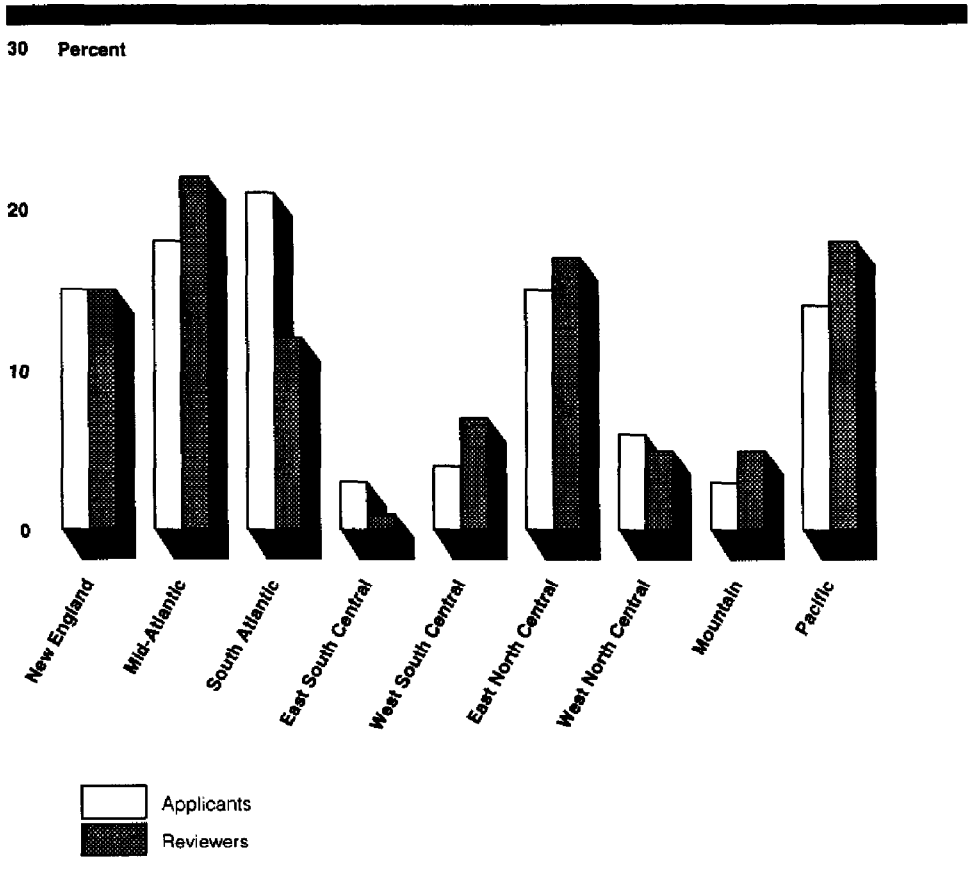
[17]We used the standard Bureau of the Census definitions for our nine regions: New England (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont); Mid-Atlantic (New Jersey, New York, and Pennsylvania); South Atlantic (Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, and West Virginia); East South Central (Alabama, Kentucky, Mississippi, and Tennessee); West South Central (Arkansas, Louisiana, Oklahoma, and Texas); East North Central (Illinois, Indiana, Michigan, Ohio, and Wisconsin); West North Central (Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota); Mountain (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming); Pacific (Alaska, California, Hawaii, Oregon, and Washington).

Figure 2.2: NIH Applicants and Reviewers by Region[a]



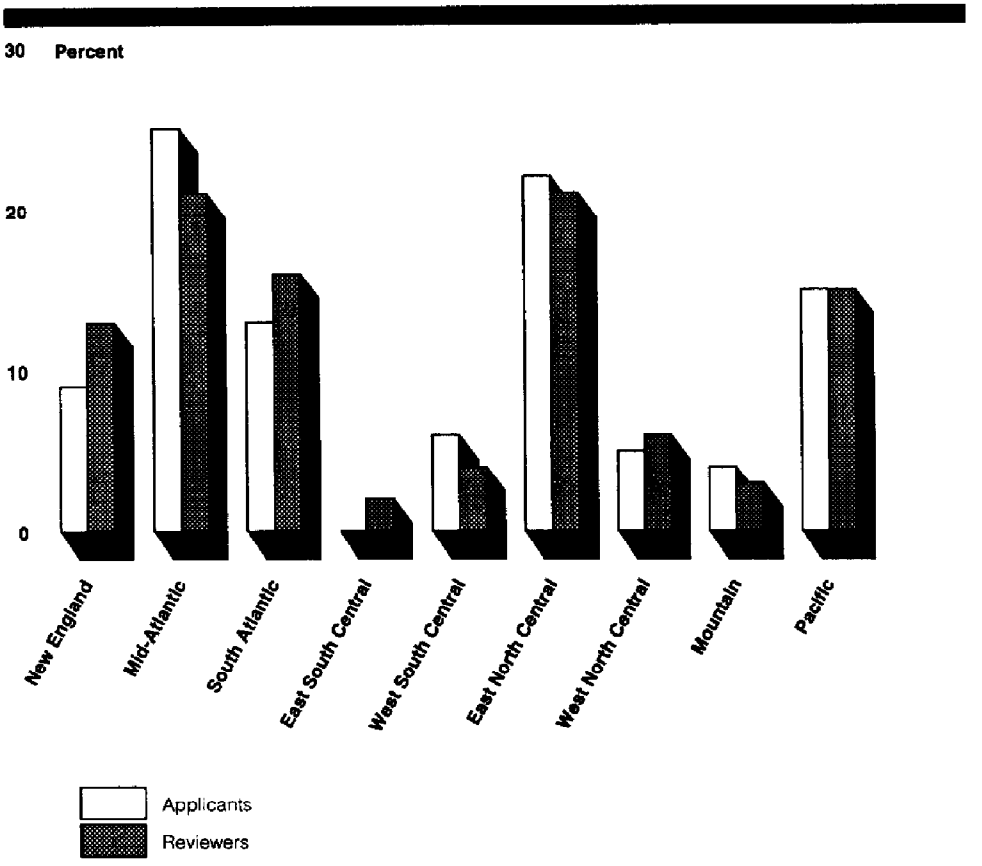
[a]Sample sizes are applicants = 97, reviewers = 589.

Figure 2.3: NSF Applicants and Reviewers by Region[a]



[a]Sample sizes are applicants = 95, reviewers = 287.

Figure 2.4: NEH Applicants and Reviewers by Region[a]



[a]Sample sizes are applicants = 315, reviewers = 468.

Note that even if the distribution of reviewers perfectly reflected that of applicants, much of the variance across regions would remain because some regions produce more applicants than others. The distribution at NEH in figure 2.4 must be seen not only in the context of reviewer representativeness but also in terms of the small number of applicants from some regions.

Contrary to critics' stereotypes, New England finished no higher than fourth among the nine regions in reviewers at any of these agencies. Rather, it was the regions known for their large research universities, the Mid-Atlantic, East North Central, and Pacific, that generally had the highest number of reviewers.

Age Representativeness of Reviewers

Next, we compared the applicants and reviewers by age distribution. We did not have the same age data across all three agencies, so we used two measures, chronological age for NIH, "professional" age for NSF, and both for NEH (tables 2.4 and 2.5). By "professional" age, we mean the number of years since the individual earned his or her Ph.D.

Table 2.4: Chronological Age of NIH Applicants and Reviewers and NEH Reviewers[a]

Agency	35 years old or younger	36-45	46-55	56 and older
NIH				
Applicants	7%	47%	32%	13%
Reviewers	0	44	47	9
NEH				
Reviewers	1	24	42	32

[a]Sample sizes are NIH applicants = 16,262, NIH reviewers = 599, NEH reviewers = 256.

Table 2.5: Professional Age of NSF and NEH Applicants and Reviewers[a]

	Year Ph.D. earned				
Agency	1980 or after	1970-79	1960-69	1950-59	Before 1950
NSF					
Applicants	42%	33%	20%	4%	1%
Reviewers	26	38	25	10	2
NEH					
Applicants	30	42	22	6	0
Reviewers	14	40	29	12	5

[a]Sample sizes are NSF applicants = 76, NSF reviewers = 277, NEH applicants = 306, NEH reviewers = 257.

For our NIH sample, more applicants were younger and older than reviewers, as shown in table 2.4. Reviewers were almost entirely in the middle of their careers, although applicants were somewhat more evenly distributed. The number of applicants peaked earlier, in the 36-to-45 range.[18]

The professional age distributions for NSF and NEH can be seen in table 2.5. At both agencies, the pattern for reviewers generally followed the distribution of applicants. The exception was for the most recent graduates, those completing their Ph.D.s in 1980 or later. This group was greatly underrepresented at both agencies.

[18]NIH applicant age data are based on aggregate fiscal year 1990 agency data, because NIH retains applicant age only in aggregate form.

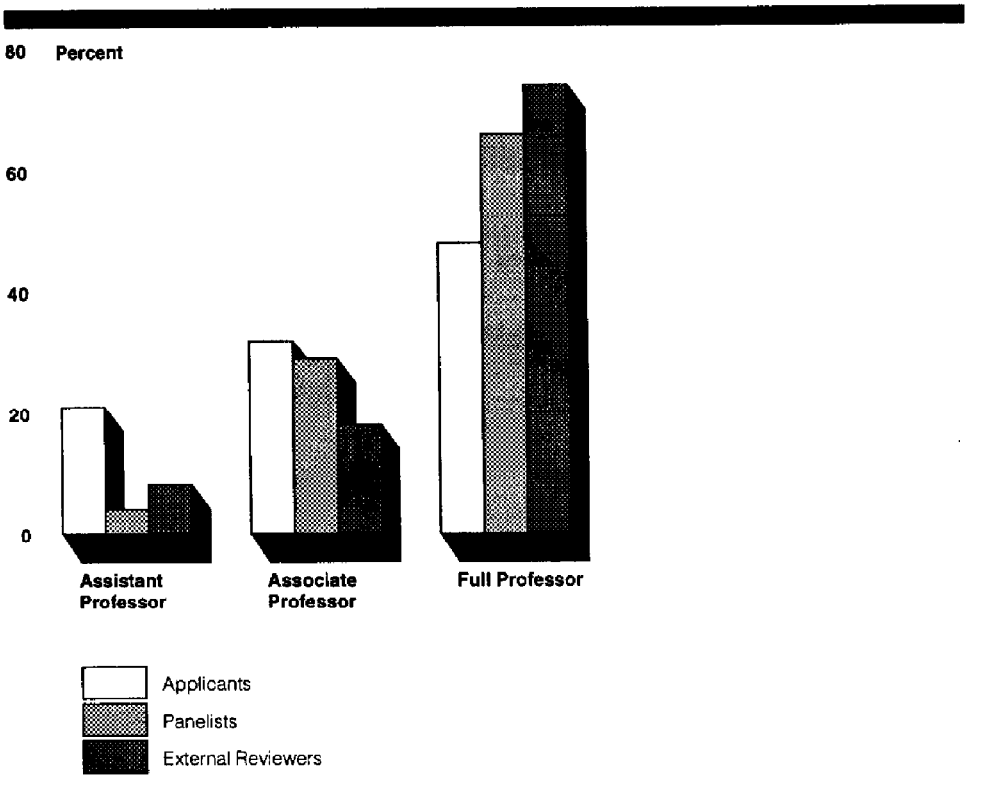
The data available to us did permit some limited comparisons across agencies. NIH reviewers tended to be younger than those selected at NEH. Almost half the reviewers at NIH were 36 to 45 years old compared to 24 percent at NEH, and NEH had four times the proportion of reviewers who were 56 or older. (See table 2.4.)

We were also able to compare the professional ages of peer reviewers at NSF and NEH. Table 2.5 shows that NEH had more reviewers in the senior categories than did NSF. Thus, NEH had older, more senior reviewers than either NIH or NSF. However, NEH consistently had more senior applicants than NSF. So the relatively older reviewers at NEH are, in part, a reflection of the agency's older applicant base. Overall, the chief concern emerging from our data with regard to reviewer age representativeness was the underrepresentation of young scholars. This pattern is consistent with that for academic rank (or seniority), to which we now turn.

Representativeness of Reviewers by Academic Rank

Only NEH had both applicant and reviewer data on academic rank: figure 2.5 summarizes the distribution for NEH participants. Applicants were twice as likely as reviewers to be assistant professors, and far fewer of them were full professors.

Figure 2.5: Academic Rank of NEH Applicants and Reviewers[a]



[a]Sample sizes are applicants = 257, panelists = 44, external reviewers = 247.

Data on academic rank for peer reviewers were available at NIH and NEH, and we acquired the rankings of NSF reviewers through our questionnaire. The great majority of panelists and external mail reviewers were full professors, and there were strikingly few assistant professors (table 2.6). NSF and NEH panelists were somewhat more likely than those agencies' external reviewers to be associate professors. That assistant professors are more commonly found among external reviewers may signify that program officers "test" potential panelists by first asking them for mail reviews.

Table 2.6: Academic Rank of Peer Reviewers at NIH, NSF, and NEH[a]

Agency	Assistant professor	Associate professor	Full professor
NIH			
Panelists	2%	29%	69%
NSF			
Panelists	0	22	78
External reviewers	5	16	78
NEH			
Panelists	4	30	66
External reviewers	8	18	74

[a]Sample sizes are NIH panelists = 460, NSF panelists = 46, NSF external reviewers = 190, NEH panelists = 44, NEH external reviewers = 247.

The underrepresentation of junior faculty, by age and rank, may reflect difficulty in identifying young scholars to serve as peer reviewers, and it is not surprising that senior researchers should more often be perceived as expert. However, the underrepresentation of younger scholars is clear from our data, and some may regard it as inherently unfair to younger applicants. Others have argued that young, junior faculty should be, and often are, discouraged by their departments from serving as reviewers because reviewer service typically is not considered to be important in tenure decisions.[19] In addition, NIH documents explicitly say "the nomination of individuals at the assistant professor level is not encouraged."[20] (NSF and NEH have no formal policy of discouraging the recruitment of junior faculty.)

Of course, the argument above can be turned around: the failure to recognize reviewer service is effectively adding a barrier to full participation by younger faculty, and, to the extent that minorities are disproportionately young, it is also a barrier to minority participation. Greater recognition of reviewer service might help remove the disincentive to serve faced by junior faculty. Making the point more broadly, the 1989 report of NIH's ad hoc panel on peer review said, "Some research institutions do little to encourage participation of faculty

[19]Marjorie A. Olmstead argued the case for why department chairs should discourage junior faculty service in "Mentoring New Faculty: Advice to Department Chairs," Committee on the Status of Women Gazette, August 1993, pp. 1-8. Jim Savage, University of Virginia, reported in a letter to GAO about an informal survey of his department chairs: "Today there was a meeting of all department chairs here at the University of Virginia, and I asked them if they encouraged or discouraged this type of participation. All of them, including women and minority chairs, said they discouraged junior faculty from such activities, as these activities did not contribute in any significant way to the eventual tenure decision."

[20]NIH, "Considerations in Nominating Study Section Members and Chairpersons," excerpted from "Nominating Members for DRG Chartered Study Sections," draft, November 1992, Bethesda, Md., p. 4.

members in study section activities, even though they are beneficiaries of research support from the NIH."[21]

A related issue is the often-stated assumption that expertise and high-quality reviews are directly related to age and experience. However, one study of review quality for peer-reviewed journals found that younger reviewers actually gave higher-quality reviews than older reviewers.[22] Another study found that lower-status reviewers give better reviews than higher-status scholars and suggested that younger reviewers find time to be more thorough.[23] In any case, these findings suggest that the underrepresentation of junior faculty should not be dismissed out of hand as an inevitable by-product of the need for expertise.

Representativeness of Reviewers by Prestige of Academic Affiliation

Some critics have argued that reviewers from prestigious universities may be preferred because of either a conscious weighing of institutional affiliation as one indicator of a scholar's merit or an unconscious prejudice in which decisionmakers unintentionally favor scholars with the "better" affiliations. So our next question is, "Are scholars from elite universities overrepresented on peer review panels?"

To examine institutional prestige, we used NAS's An Assessment of Research-Doctorate Programs in the United States to rank all our

[21]NIH, Sustaining the Quality of Peer Review: A Report of the Ad Hoc Panel (Bethesda, Md.: NIH, 1989) p. 4. When she was NSF's senior science adviser, Mary Clutter made this same point, telling a congressional hearing that "one of the reasons reviewers from undergraduate institutions are reluctant to serve is . . . this valuable voluntary service to the government and the community is too rarely recognized by their own institution." See Hearings on Research Project Selection, Task Force on Science Policy, Committee on Science and Technology, House of Representatives (Washington, D.C.: April 9, 1986). Former NSF Director Erich Bloch also made the point in a letter to the president of Oberlin College made available to the committee. Surveys of reviewer attitudes, including our own survey, have found that reviewers believe greater recognition and remuneration would promote better, more careful reviews. The current, virtually pro bono service forces participants to take time away from income-generating activities. See Chubin and Hackett, p. 205.

[22]Arthur Evans et al., "Characteristics of Peer Reviewers Who Produce Good Reviews," presented at the Second International Congress on Peer Review in Biomedical Publication, American Medical Association, Chicago, Ill., September 9, 1993.

[23]Thomas Stossel, "Reviewer Status and Review Quality," The New England Journal of Medicine, March 7, 1985, pp. 658-59. Stossel adds several alternative explanations: "Lack of time may be a legitimate excuse for refusing to serve but not for poor reviewing. Some senior scientists may be too removed from the front lines of research to have the grasp required to criticize effectively. Alternatively, high status referees may rely excessively on personal authority. . . . Seen in this light, lack of time to review papers may in some cases be more a habit than a reality" (p. 659).

university applicants and peer reviewers.[24] However, the NAS study did not rank university-affiliated professional schools such as schools of medicine, dentistry, or pharmacy. To rank these, we used the Gourman report.[25] The two studies use somewhat different criteria and methodologies but are similar enough for our purposes in providing an empirical, survey-based ranking of institutions.

The NAS study ranked departments on a variety of criteria. We used the NAS ranking of departments by the quality of their faculty. The top departments had scores in the 30-39 range, average departments about 50, and lower-ranked departments had scores in the 70s.

Table 2.7 summarizes NAS rankings of applicants and reviewers from university departments. At NIH and NSF, more applicants than reviewers were from higher-ranked schools—that is, those rated in the top two categories. Conversely, reviewers from lower-ranked schools were overrepresented when compared to their applicant peer group. At NEH, there were some disparities between prestige categories but not the trends we saw at NIH and NSF. Top schools were slightly overrepresented at NEH, but the 7 percent of reviewers in this category still included fewer reviewers than in any other category.

[24]Lyle V. Jones, An Assessment of Research-Doctorate Programs in the United States, vols. 1-5 (Washington, D.C.: National Academy of Sciences, 1982). Used for the following fields: art history, biochemistry, botany, cellular and molecular biology, chemistry, classics, economics, electrical engineering, geography, linguistics, mathematics, microbiology, philosophy, physics, zoology, and language and literature in English, French, German, and Spanish. We used this ranking because, although dated, it was the most thorough and comprehensive.

[25]Jack Gourman, The Gourman Report (Los Angeles, Calif.: National Education Standards, 1989). We used this for medical schools, pharmacy schools, and programs in veterinary medicine and sciences.

Table 2.7: NAS Prestige Rankings for University Departments of NIH, NSF, and NEH Applicants and Reviewers[a]

Agency	Top ranked: under 40	40-49	50-59	60-69	Lower ranked: greater than 70
NIH					
Applicants	11%	28%	44%	17%	0
Reviewers	4	25	32	35	4%
NSF					
Applicants	3	35	35	16	12
Reviewers	1	20	38	29	12
NEH					
Applicants	0	25	21	54	0
Reviewers	7	17	25	38	14

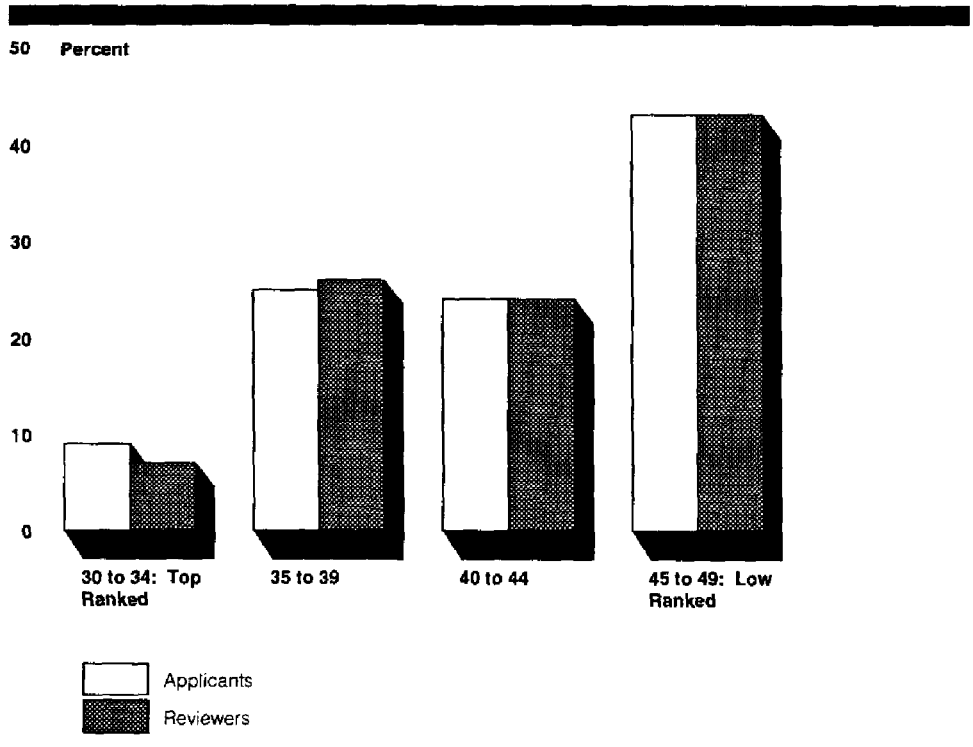
[a]Sample sizes are NIH applicants = 18, NIH reviewers = 97, NSF applicants = 69, NSF reviewers = 169, NEH applicants = 24, NEH reviewers = 96.

Source: Lyle V. Jones, An Assessment of Research-Doctorate Programs in the United States, vols. 1-5 (Washington, D.C.: National Academy of Sciences, 1982).

At NIH, some of the participants come from traditional university arts and sciences departments ranked by NAS, but most are employed at professional schools, which are rated by the Gourman study. Figure 2.6 shows that panelists in our sample employed at a school of medicine, pharmacy, or veterinary medicine were highly representative of sample applicants. Moreover, the proportion of both applicants and peer reviewers tended to decrease as department prestige increased. Although the patterns for medical school and university department participants were quite different, the common end-result for NIH was that elite schools were not overrepresented.[26]

[26]The low number of applicants in the top categories of the Gourman rankings may be an artifact of that method or of our collapsing rankings to these category breaks.

Figure 2.6: Prestige Rankings for Professional Schools of NIH Applicants and Reviewers[a]



[a]Includes medical, pharmacy, and veterinary schools. Sample sizes are applicants = 1,023, reviewers = 165.

Source: NIH data and Jack Gourman, The Gourman Report (Los Angeles, Calif.: National Education Standards, 1989).

However, the majority of university departments were not ranked at all. There are, after all, thousands of colleges and universities in the United States, and the rankings rarely rate more than 100. We looked at the distribution of reviewers and applicants whose departments NAS did not rank. At NSF and NEH, unranked reviewers slightly exceeded the proportion of unranked applicants. So at these agencies, reviewers from unranked departments appear to be included representatively. However, at NIH, unranked schools were underrepresented among reviewers.[27] This contrast may stem from the greater use of ad hoc panels and external reviews at NSF and NEH, which means they recruit more reviewers and therefore draw from a larger proportion of persons active in their fields than NIH, which uses primarily standing panels.

[27]All the unranked data were based on the NAS ranking of departments. The Gourman report ranks a large number of medical schools, so there are relatively few, if any, unranked medical schools.

Overall, reviewer prestige was roughly representative of applicant prestige. In fact, across the agencies, reviewers were less likely to come from prestigious universities than were applicants. Unranked departments were well represented except at NIH. The evidence generally contradicts the allegations of bias in reviewer selection on the basis of departmental prestige. These data also support the argument of program officers that top reviewers are hard to recruit. This could stem from several possible factors. Top scholars may be busier, or more likely to have funding relationships with other research grant programs, than other scholars. It may also be the case, as Jerome Green, Director of the NIH Division of Research Grants, has noted, that they used to get "the greatest names," but now "most 'top echelon' scientists have served, many feel once is enough."[28]

Summary and Conclusions

We examined the selection of reviewers regarding their expertise and several dimensions of how well they represent the agencies' applicant base. Overall, the agencies did quite well on some factors, had a mixed record on others, and were fairly unrepresentative on some dimensions.

Reviewers at all three agencies generally represented the prestige and regional variety of applicants. We found little evidence to support the notion that peer review panels are staffed disproportionately from among researchers at a handful of elite institutions. On the contrary, among our sample cases those from elite institutions tended to be somewhat underrepresented among reviewers compared to applicants. Although we found notable disparities in the number of reviewers across regions, this was not an issue of selection preference or representativeness with regard to applicants. Instead, it showed differences in the underlying applicant base, reflecting other questions such as regional distribution of talent and resources and agency outreach to applicants.

The agencies had more of a mixed record on gender, expertise, and personal familiarity between the participants. We found that although women served as panelists, at least in proportion to their representation among sample applicants at all three agencies, they were underrepresented among external reviewers in some programs at NSF. This is especially important because of NSF's heavy, and in some disciplines exclusive, reliance on outside reviews.

[28]"Conflicting Agendas Shape NIH," Science, 261 (September 24, 1993), 1678.

In general, we found that peer reviewers at all three agencies reported some research expertise in the general area of, or an area related to that of, the proposals they reviewed but that a majority of reviewers at NIH and NEH, and a near majority at NSF, could cite only a few or no references to the relevant literature.

We also concluded that there was some tension between the ideal of a high level of reviewer expertise on the topic of a research proposal and a low level of personal familiarity with the applicant. That is, the greater the expertise that all agree is needed in a reviewer, the greater the possibility of personal knowledge and the potential for either positive or negative bias. NIH reviewers reported the least proximate research expertise, and its panelists also indicated less personal familiarity with the applicants than the reviewers at NSF but more than those at NEH. In contrast, NSF reviewers indicated the most proximate research interests, and they were also the most likely to know the applicants. As we noted earlier, NSF tends to be organized along disciplinary lines to a much greater extent than either NIH or NEH. This organization, and the fact that some NSF programs select panelists and outside reviewers in part to reflect the topics of current proposals, helps account for the closer affinity of research interests and the greater degree of personal familiarity between reviewers and applicants reported by our NSF respondents. However, this situation could be exacerbated by the practice in some cases of allowing panelists to select proposals for which they would serve as primary or secondary reviewers and the policy in some programs, as discussed above, of permitting applicants to suggest persons who should (or should not) review their proposals.

Finally, young researchers and those with lower academic rank tended to be underrepresented at all three agencies, despite efforts to ensure outreach to these groups. This may stem in part from difficulty in identifying younger, lower-ranked researchers who have not yet established visibility in the field and efforts by institutions to discourage their participation in peer review.

Rating of Proposals by Peer Reviewers

Neither the best intentions of administrators nor the worst problems in the selection of peer reviewers may actually determine or influence how reviewers behave once they have a proposal before them. In other words, any evidence of unrepresentativeness among reviewers raises the question of potential bias but does not prove that biased behavior actually occurred. So the next step is to examine the behavior of reviewers. We cannot answer what causes reviewers to give the ratings or scores they do; however, we can and do address the question of which factors are at least empirically related to how reviewers rate proposals and which are not.

A grant proposal must meet a variety of criteria set by the agencies. However, none of the agencies currently rate how proposals meet each of their criteria. Instead, they only give each proposal a single summary rating. For example, none of the agencies gives a separate rating for the contents of the proposals —that is, the importance of the topic and the quality of the design.[1] Therefore, neither we nor the agencies can directly measure the weight that is given to these elements apart from considering other factors, such as the track record of the applicant, the strength of the applicant's institution, or the reviewer's personal knowledge of the applicant. We can only infer how much questions and designs count by considering how much of the variation in ratings remains unexplained after all the other factors have been considered.

This also means that any relationships we find in our data between a given factor and rating may be spurious. So for most of the findings in this chapter, rival hypotheses can be entertained—that the relationship to the rating, or score, reflects bias based on some characteristic of the applicant or stems from real differences in the quality of the proposals submitted by scholars who differ on that characteristic. For example, if we find that men get higher scores than women, we cannot empirically resolve whether this is because of reviewer bias against women or because women are submitting weaker proposals.

Both internal and independent evaluators who have tried to study peer review have complained for some time about the unavailability of data. In 1974, the NAS Committee on Science and Public Policy was "concerned about the lack of good evidence in the government-wide peer review

[1]As one NIH internal report put it: "At present no formal weighting system is given to grant review criteria, and the general impression is that their relative importance varies among applications and among study section members." Report of the NIH Peer Review Committee (Bethesda, Md.: December 1988), p. 8.

process."[2] A recent history of NSF characterized testimony before the Congress in 1975 by the agency's top evaluator as saying "it was unfortunate the NSF had not been collecting data of this sort on a systematic basis over a longer period of time. . . . [I]t would have established that no bias existed against some region or school."[3] Fifteen years later, scholars were still frustrated by the inadequacy and inaccessibility of agency peer review data.[4] In conducting this study, we found that, in contrast to NSF, NIH did not retain records of individual reviewer scoring. NIH discarded race identification data, and NEH did not collect them. Yet these data are critical to evaluation efforts for congressional oversight.

Agency Procedures and Criteria

Before we examine how various factors were related to the scoring of proposals by our sample respondents, we review each agency's procedures and guidelines related to scoring.

NIH Review Policies

All NIH research grant applications (R01s) are assigned to an initial review group for peer review. Each application is assigned two or three primary reviewers, who present to the rest of the panel a summary of the proposal and their analysis of the proposal's merits. If there are very serious problems with the proposal, there is a motion to "not recommend for further consideration" and, if the motion passes, no further review is conducted and no rating is given.

Of course, the overwhelming majority of proposals are given full reviews and a scientific merit rating by each primary reviewer.[5] At the end of the discussion of each proposal, it is given a rating by each of the panelists; then the average of these ratings is multiplied by 100 to get the priority score.

To reduce the influence of disparities in scoring between meetings of the review panel, the priority scores are pooled with the scores of proposals in the previous two meetings of the panel and given a percentile score relative to this larger base. The percentile score is the ranking considered by the institutes in their funding decisions.

[2]George T. Mazuzan, "Good Science Gets Funded," Knowledge, September 1992, p. 81.

[3]Mazuzan, p. 79.

[4]Daryl E. Chubin and Edward J. Hackett, Peerless Science: Peer Review in U.S. Science Policy (Albany, N.Y.: SUNY Press, 1990).

[5]NIH has recently announced an experiment to "triage" applications—that is, to sort out the weaker ones before they get a full review.

The review criteria include the significance and originality of the proposal from a scientific and technical standpoint; the adequacy of the methodology; the qualifications and experience of the principal investigator and staff; the availability of institutional resources and reasonableness of the proposed budget and study duration; and human subject, animal welfare, and biohazard factors.

NSF Review Policies

Each program at NSF separately decides whether to use a numerical or alphabetical rating system. But all peer reviewers within a particular program, whether external or panelists, apply the same criteria and rating scale to proposals. When panels are used, members volunteer to take primary or secondary review responsibility for several specific proposals. Any proposals left without primary reviewers are then assigned by the program officer to individual panelists.

The National Science Board established the criteria for the selection of research projects by NSF. These include (1) the capability of the applicant and project team, the proposal's technical soundness, and the adequacy of institutional resources and (2) the likelihood that the study will advance basic or applied research or the infrastructure of science and engineering. Infrastructure development includes the education and participation of women and minorities, expanding opportunities across geographic areas, and assisting underdeveloped fields and new investigators.[6]

NEH Review Policies

At NEH, panelists are expected to have read all the proposals and to have given them a written review and preliminary rating before the panels meet. After discussion by the panel, each member assigns a final rating. The program officer writes a summary of the review and compiles a ranked list of proposals that he or she will recommend for funding.

At NEH, the evaluation criteria vary somewhat by program, but the common threads are the applicant's qualifications to perform the proposed work, based on his or her past work; the significance of the project to the humanities; the clarity and soundness of the proposal; the likelihood that the applicant will complete the project; and the adequacy of institutional support.

[6]NSF, Grants for Research and Education in Science and Engineering (GRESE) (Washington, D.C.: 1990). An updated version was published in January 1994.

Common Problems in Reviewer Education Procedures

Accuracy and consistency in scoring depend not only on the reviewers' command of their discipline but also on their knowledge of the agency's review criteria.[7] All the agencies do little to ensure that reviewers have an accurate and similar understanding of the agency's criteria and rating scales. Nor are reviewers advised about the relative weight to be given to an agency's criteria. This is likely to be more of a problem at NSF and NEH, less so at NIH, because only NIH consistently uses review by standing panels, which allows reviewers to become quite familiar with the criteria and acquire a sense of how their peers are applying them. It is still a problem at NIH as evidenced by replies to our open-ended questions from NIH reviewers who noted the inability of the current scoring system to effectively sort out the best proposals from the merely very good ones. However, NEH reviewers were more concerned about criteria being selectively employed.

The Problem of Unwritten Criteria and Insider Knowledge

A problem we identified in our observations of panels at all the agencies was that reviewers often relied on unwritten or informal criteria when evaluating proposals. The most common of these was an expectation that a proposal include some preliminary results that would demonstrate the feasibility of completing the project.

This criterion of preliminary results was widely cited and applied at NIH, and NIH officials have acknowledged that preliminary results are an important criterion in grant review.[8] While instructions for NIH applications indicate that a discussion of preliminary results is optional, we observed that it was essential. One study section member has remarked that "Preliminary data in a new grant application are about as optional as breathing."[9] The existence of these unwritten criteria was also confirmed by agency officials at NSF. NEH reviewers whom we observed frequently said they had downgraded proposals that failed to adequately develop the underlying philosophy of the work or failed to consider gender issues or failed to use "the right translation" or, most frequently, failed for "lack of sufficient preliminary research." While these are arguably important matters to consider, if reviewers regularly take them into account, then applicants should know about them.

[7]See "Calibration," in Michael Scriven, Evaluation Thesaurus, 4th rev. ed. (Newbury Park, Calif.: Sage, 1991).

[8]See John McGowan, "NIH Peer Review Must Change," The Journal of NIH Research, 4 (August 1992), 58.

[9]Quoted from Liane Reif-Lehrer, The Grant Application-Writers Handbook, forthcoming.

Although it was impossible to relate such criteria to the ratings on reviews statistically, the existence of unwritten criteria does raise concerns about fairness, because it tends to work to the disadvantage of persons not previously part of the agency's funding system. Presumably, those who have served as reviewers and recipients of prior grants would know about the unwritten rules from experience, but newer applicants could not be expected to have such information. Unwritten decision rules create the opportunity for insider knowledge. Particularly where feedback mechanisms are weak, this could bias the outcomes of the peer review process. In the extreme, unwritten criteria could create a privileged class of scholars who, by having served as reviewers, would have gained an advantage in future grants competition.

Such an advantage would, of course, be unfair, and it is especially important to ensure that the situation does not arise in light of studies showing that applicants who have been reviewers or advisory council members for an agency to which they are applying are much more likely to get funded than those who have never reviewed for that agency.[10] One study of peer review at NIH found that both study section members and advisory council members, past and present, were more likely to have their research grant applications approved. The study section members got much better ratings than nonmembers, suggesting they were submitting better proposals. But there are several reasons why reviewers may get better ratings: (1) since they are chosen for their expertise, they may be better researchers and better at writing proposals; (2) they may have profited from knowledge they gained of the state of the disciplinary research agenda while sitting on panels; (3) their knowledge of panel decisionmaking, and especially of unwritten decision rules, may have helped them write more highly rated proposals; (4) it could be that panel members were biased in their favor.[11] Similarly, an internal NSF study found that 75 percent of applicants had served as reviewers but that of the applicants who were consistently successful, 97 percent were reviewers or had served as reviewers in the last 5 years.[12] However, a study by the NIH peer review committee found that while study section panel members

[10]Since, as we reported in chapter 2, junior faculty are underrepresented among reviewers, they may be particularly disadvantaged by unwritten decision rules.

[11]See Adil Shamoo, "Role of Conflict of Interest in Public Advisory Councils," Ethical Issues in Research, FIDA Research Foundation Proceedings (Frederick, Md.: 1992). NIH staff performed the statistical analysis for this study.

[12]NSF, Program Evaluation Staff, Proposal Review at NSF: Perceptions of Principal Investigators (Washington, D.C.: February 1988).

consistently rated better than nonmembers, their ratings did not increase during membership.[13]

Further, controversy exists about the unwritten criterion of preliminary results in its own right, because it may shift the basis of awards from proposals of future work to a retrospective recognition of accomplished research. Currently, all agencies do apply retrospective considerations, such as the applicant's track record, in assessing the merit of proposed future research. The criterion of preliminary results goes a step further by shifting the process to a need for, and recognition of, past research. Since already accomplished performance can be evaluated with more certainty than future success, the criterion of preliminary results is a way of reducing the risk of funding bad research. However, it raises barriers to newcomers and may erode the legitimate risk-taking behavior of reviewers needed to fund ambitious studies with important potential.[14]

Peer Review in Practice: Factors Related to Reviewers' Scoring

It has often been alleged that criteria other than those outlined in official policies—an applicant's gender, for example—influence reviewers' scoring. We used our sample of survey respondents and the corresponding agency administrative data on reviewers and proposals to examine the empirical basis for some of these concerns. Our analysis question was, What factors influence ratings? First, we consider a number of factors separately, and then we examine all the factors together and discuss which relationships continue to hold.

Because the agencies use different scoring procedures, we have recalculated ratings so that results can be viewed comparatively. To do so, we converted all ratings to a 5-point scale, on which 1 is the best and 5 the worst score. Thus, the lower the numerical score, the better the rating of the proposal.

Ratings and Reviewers' Expertise

One complaint made by applicants for federal grants is that reviewers often lack the relevant expertise to fairly judge their proposals. The result, they allege, is that reviewers, unable to understand the merits of the proposal, rate them too harshly. Relatedly, some claim that reviewers are inclined to give better ratings to proposals in their own particular

[13]NIH, Report of the NIH Peer Review Committee (Bethesda, Md.: December 1988), p. 32.

[14]John McGowan, Director of the Division of Extramural Activities, of the National Institute of Allergy and Infectious Diseases, has argued that this doctrine of preliminary results "discourages the submission or favorable review of research grants that are innovative or high-risk." See McGowan, p. 58.

subfields, seeking to strengthen the claims on federal funds of those who share their interests. Others contend that, on the contrary, reviewers tend to regard applicants in their own areas as competitors and, therefore, give these proposals worse ratings.

As we saw in chapter 2, for the most part reviewers were working in areas generally or closely related to the topics of the proposals they reviewed, although seldom on the same question. We examined the relationship both of reviewers' self-reported proximity of research interests to the specific proposals they reviewed and of their knowledge of the relevant literature to their ratings of those proposals. Tables 3.1 and 3.2 report the results of this analysis. Each cell of the tables shows the mean rating given by reviewers with varying levels of research proximity to the proposals.[15] Table 3.1 shows that at NIH reviewers with the least proximate interests gave the proposals they reviewed the highest (worst) ratings, on the average. However, at both NSF and NEH, the reviewers whose interests and research were the least proximate gave better ratings, on the average. Table 3.2 reports how ability to cite the literature relevant to the proposal, which is another measure of research proximity, was related to the rating score. The patterns seen in table 3.1 were confirmed for NSF but not for NIH and NEH.

Table 3.1: Mean Rating Scores by Reviewers' Research Proximity to the Proposal[a]

Agency	Same or related area	General area	Unrelated area
NIH	1.9	2.2	2.2
NSF	2.1	2.1	1.7
NEH	2.9	2.4	2.4

[a]Sample sizes are NIH = 109, NSF = 157, NEH = 72.

Table 3.2: Mean Rating Score by Reviewers' Ability to Cite Literature[a]

Agency	Could cite many references	Could cite only a few references	Could not cite references
NIH	2.3	2.0	2.3
NSF	2.1	2.1	1.5
NEH	2.7	2.3	2.5

[a]Sample sizes are NIH = 106, NSF = 158, NEH = 73.

[15]The number of cases for each agency is lower than the total number surveyed. For NIH and NSF, this reflects oversampling, as discussed in chapter 1; for NEH, the tables reflect only panelists because mail reviewers there did not give scores.

Regarding NSF, these findings could mean either that those with the most relevant expertise were able to detect more flaws in the proposed work or that reviewers were downgrading competitors. In either event, there is little evidence to support assertions that reviewers want to promote their fields and therefore give preferential ratings to proposals in their own immediate subfields.

Rating Scores and "Matthew" or "Halo" Effects

Another set of concerns centers on the perceptions of applicants and their institutions that reviewers bring to the task of scoring proposals. Many argue that ratings often reflect not so much an evaluation of a specific proposal up for review as a set of expectations reflecting the prior work or "track record" of the individual or the reputation of the institution for which the applicant works. In other words, proposals from highly regarded scholars, or from those at highly regarded academic departments, are likely to be given the benefit of the doubt when the proposals are reviewed, whereas proposals from unknown applicants or from departments without strong reputations are not. Recall that reviewers at all three agencies are supposed to take account of the competence of the applicant and likely institutional support, so it is basically proper for reviewers to consider these factors. However, reviewers can reduce competence to reputation and publication record, just as the agency criterion regarding "adequacy of institutional support" can become a euphemism for affiliation with a prestigious institution.

First, we look at the relationship between a reviewer's perception of the applicant's track record and ratings—what is called the "Matthew effect." The Matthew effect is a term coined by sociologist Robert Merton from the biblical book of Matthew: "For unto everyone that hath shall be given, and he shall have abundance; but from him that hath not shall be taken away even that which he hath."[16] In other words, scholars who have in the past been successful will be judged as more likely to succeed in the future and therefore receive better peer review ratings than scholars with otherwise equivalent proposals but weaker track records.

For our purposes, the importance of the applicant's track record lies less in the actual number of publications (or citations of those publications by others) and more in how the reviewer perceives the applicant's track record. Therefore, we asked each of our survey respondents to rate the applicants whose proposals they reviewed as major figures in the field, notable contributors, competent researchers, or unknown researchers.

[16]Matthew 13:12; see Robert K. Merton, "The Matthew Effect in Science, II," ISIS, 79 (1988), 608-9.

The results shown in table 3.3 are clear and consistent across all three agencies: the better the applicant's perceived track record, the better the rating, on the average.[17]

Table 3.3: Reviewers' View of Applicants' Track Record and Mean Rating Score[a]

Agency	Major figure	Notable contributor	Competent researcher	Unknown researcher
NIH	1.6	1.9	2.4	3.3
NSF	1.6	1.8	2.6	3.4
NEH	1.8	2.2	2.7	3.7

[a]Sample sizes are NIH = 99, NSF = 157, NEH = 61.

Next, we consider the halo effect. A halo effect occurs when a proposal from an applicant is viewed more favorably because he or she is from a prestigious university department. Being associated with distinguished colleagues can give an advantage to an applicant in two ways. First, even if the applicant is not well known, reviewers may infer that he or she is likely to use the grant productively just because a prestigious department has demonstrated its own confidence by hiring the individual. Second, the reviewers may also be counting on the individual's colleagues to provide the assistance and advice that make success more likely. Indeed, at the panels we attended there were several occasions when the reviewers discussed the level of help likely from colleagues; usually these were references about co-investigators, but sometimes the discussions turned to colleagues not listed. In short, just being part of an elite academic department may cast a halo over the applicant that leads to a more positive evaluation of the proposal than might otherwise have been made.

To examine this possibility, we used survey data on the reviewer's perception of the applicant's academic department. Respondents were asked to rate the department as among the top 5 in the discipline, among the remainder of the top 20, or not in the top 20. The results are reported in table 3.4. Once again, the pattern is clear and consistent across all three agencies: the more prestigious the applicant's department is perceived to be by the reviewer, the better the rating.

[17]Of course, the causal arrow could run the other way. Reviewers could be using the score they gave to help rate retrospectively the applicant's track record. However, many of our respondents reported keeping extensive records of their reviews and the proposals, making this outcome unlikely.

Table 3.4: Reviewers' View of Applicants' Department and Mean Score[a]

Agency	Rated among top 5	Rated among 6th-20th	Not rated among top 20
NIH	1.5	2.0	2.3
NSF	1.6	1.9	2.4
NEH	1.7	2.3	3.0

[a]Sample sizes are NIH = 98, NSF = 154, NEH = 57.

Scores, Conflict of Interest, and Personal Familiarity With Applicants

Another concern frequently voiced about peer review is that reviewers may be inclined to use their positions to help promote the projects of friends and colleagues (or, conversely, to hurt the chances of rivals). All three agencies have conflict-of-interest policies designed to reduce this threat to some degree. In general, reviewers are asked not to judge proposals from their employing institutions or those for which they are listed as participants or consultants. Furthermore, panelists are required to absent themselves from discussions of such proposals.

However, in observing fiscal year 1993 panels, we noted that, while these provisions usually were well-enforced, there were lapses in a number of instances at each agency.[18] Moreover, conflict-of-interest rules do not necessarily disqualify panelists, and outside reviewers often have personal relationships with applicants other than employment at the same institution or participation on the same grant proposal. These relationships are difficult to measure and can take many forms (such as personal friendship, past collaboration, teacher-pupil status). Even if formal conflict-of-interest problems are avoided, it is possible that reviewers could help give advantage (or disadvantage) to proposals from those they know personally. As we saw in chapter 2, most reviewers at all three agencies had at least indirect knowledge of the applicants, with NSF reviewers having direct knowledge most frequently.

For each agency, we calculated average scores for the cases in which reviewers reported direct, indirect, and no personal knowledge of the applicants. The results are shown in table 3.5. At both NSF and NEH, but not

[18]For example, at one NSF panel, the final rank ordering of proposals was done with all panelists present, including several who had excused themselves from earlier discussions because of conflicts of interest. Under the circumstances, it would have been possible for a panelist to improve a colleague's chances of funding by arguing to move higher-ranking proposals down on the list relative to the colleague's. In another case, at NIH, a clerk assigned to monitor and enforce conflict-of-interest rules left before the meeting ended. Subsequently, one panelist remained in the room during a discussion of a proposal from his home institution, a clear case of conflict, as confirmed by an agency official attending the meeting. One possible contributing element to such problems is that the master list used to track institutional conflicts failed to always list the campus of the applicants.

NIH, the pattern was clear. Proposals from applicants whom reviewers knew directly got the best scores on the average, while those from applicants known only indirectly through a third party received somewhat lower ratings, and those from applicants unknown to reviewers fared worst. In contrast, at NIH there were virtually no differences.

Table 3.5: Reviewers' Familiarity With Applicants and Reviewers' Mean Scores[a]

Agency	Direct	Indirect	Neither
NIH	2.1	2.0	2.3
NSF	1.7	2.3	2.8
NEH	2.0	2.3	3.1

[a]Sample sizes are NIH = 103, NSF = 154, NEH = 57.

These results at NSF and NEH may mean, as many fear, that familiarity improves acceptance. But they may also simply reflect other relationships we have seen already. For instance, scholars with great track records may also be more likely to attend professional conferences where they can meet others in their disciplines, expanding their networks both directly and indirectly. Past success can mean more grant money for travel and publication support and can build a cumulative advantage. As we saw in table 3.3, track record is also highly related to scores.

Scores and Race

As we noted in chapter 2, we could not obtain data on the race of reviewers from the agencies. In addition, only NSF was able to provide data on the race of its applicants. For that agency, race did appear to be related to scores. On the average, applications from whites received scores of 2.0, compared to 2.8 for nonwhites. It is not obvious why this occurred, since race and ethnicity information is collected on a separate form and withheld from reviewers.[19] However, other data (applicant's name or institutional affiliation, personal knowledge of the applicant) could convey this information. In fact, in our panel observations we listened as panelists tried to sort out the gender and race of applicants. Nonwhites might be identified by the ethnic origin of their names or their affiliation with traditionally minority universities. Recall also that 86 percent of NSF reviewers had at least indirect knowledge of applicants and would often know an applicant's racial background in spite of confidentiality procedures. A major concern here is that minority applicants may be disadvantaged in peer review. However, it may also be that minority

[19]Completing and returning these forms is voluntary, and the response rate may have decreased in recent years.

researchers at NSF simply wrote proposals of lower quality. We have no evidence to conclude one way or the other.

Regression Analysis of Factors Related to Review Scores

As we noted in the previous section, a number of factors appear to be related to proposal scores. However, some of these factors may also be related to one another, so the apparent relationships with scores may be spurious. To complete our analysis, therefore, we reanalyzed these relationships, taking account of all the factors simultaneously, using multiple regression analysis. This is a statistical technique that estimates equations showing the relationship between a dependent variable (in this case, the scores received by each proposal) and a set of independent variables (here, a variety of factors, as discussed earlier). The major advantage of this technique is that it allows us to examine the relationship between the dependent variable and each independent variable while controlling for the effects of the other independent variables.[20]

To conduct this analysis, we needed to transform several of the variables into somewhat different forms. For largely categorical variables, we used dummy variables. For example, we recoded the gender variable, setting female equal to 0 and male equal to 1; thus, the estimated regression coefficient for this variable may be interpreted as the relationship between an applicant's being male and the reviewer's score.

The variables entered into this model were an applicant's perceived track record, the reviewers' ranking of the applicant's department, the applicant's gender, age or professional age, region, amount and duration of request, as well as reviewers' proximity of expertise and knowledge of the literature.[21] Applicant's race was also used where it was available at NSF.

Factors Related to Scores at NIH

Our findings for NIH, presented in table 3.6, show that most of the factors considered in the previous section drop out when considered together. The first and strongest factor represents a reviewer's perception of an applicant as a notable contributor to his or her field. Applicants who are

[20]The specific technique we employed is called stepwise regression. In this approach, we first entered all the independent variables into the equation and then allowed the computer to remove, one at a time, those not significantly contributing to the amount of variation that could be explained by the equation, until only the significant relationships were left. The result is that different equations were estimated for each agency, indicating that different factors were related to the ultimate scores. We report the results for each agency separately.

[21]For the regressions, we divided the cases into four regions rather than nine as before because of the need to create dummy variables for each region.

considered by their reviewers to be notable contributors score .56 lower (on a 1-to-5 scale) than others. Gender was the next most important factor related to score: men on the average had much better scores than women. The longer the length of time requested for the research, the better the score. Each additional year of support was related to a .26-lowering of the score. The other factors are statistically significant but have only the smallest effect on score.[22]

Table 3.6: Factors Related to Proposal Scores at NIH[a]

Independent variable	Regression coefficient	Standard error
Constant	3.50	.44
Applicant is perceived as notable contributor	-56	.16
Applicant is male	-.46	.22
Requested duration (months)	-.26	.11
Requested amount (per $10,000)	.007	.003
Requested amount (per $10,000) and principal investigator perceived as major figure	-.01	.002

[a]Adjusted R^2 = .24. Sample size = 94.

Most striking is the fact that the overall model represented by this equation explains only a small part of the variation in scores. This is reflected in the adjusted R^2 of .24, which indicates that, statistically, the model explains only 24 percent of the variance among scores. Recall that separate data regarding the importance of the question raised in the proposal and the strength of the design to address that question were unavailable. Presumably, these are the most important considerations in scoring proposals. Not surprisingly, then, our other variables do not explain much of the variance. However, the low amount of variance explained could result from the fact that the scores used in this chapter are the scores reported by the individual reviewers in our survey. These are sometimes based on memory rather than on the reviewer's records. For NSF and NEH, all score data are from agency records.[23]

[22]The request amount coefficient is for units of $10,000. A request increase of $100,000 would hurt the score only by a trivial .07. The last coefficient of -.01 can be interpreted as indicating that for any given amount requested major figures had an advantage, but the coefficient is so small as to be trivial.

[23]The NIH percentile scores for each proposal were available and are the basis for the NIH data in chapter 4, where average scores are the appropriate unit of analysis.

Factors Related to Scores at NSF

Table 3.7 shows our analysis for NSF. An applicant's track record was the most important factor related to scores. The top four rows show that, in general, the better the perceived track record, the better the scores. Indeed, being seen as either a major figure or a notable contributor is related to a whole point improvement in score over being seen as merely competent. The next two rows of table 3.7 are for gender and race. Being male or white is related to having a higher score. The reviewer's being personally familiar with or having indirect knowledge of the applicant also is associated with the application's getting a better rating, but interestingly the level of familiarity does not matter much because personal knowledge counts for little more than indirect knowledge (-.49 versus -.44).

Table 3.7: Factors Related to Proposal Scores at NSF[a]

Independent variable	Score regression coefficient	Standard error
Constant	5.4	.53
Applicant is perceived as major figure	-1.6	.38
Applicant is perceived as notable contributor	-1.4	.31
Applicant is perceived as competent researcher	-.5	.30
Applicant is male	-.67	.39
Applicant is white	-.54	.22
Reviewer is directly familiar with applicant	-.49	.26
Reviewer is indirectly familiar with applicant	-.44	.25
Applicant's professional age	.02	.008
Requested amount (per $10,000)	.001	.0004
Requested duration (months)	-.03	.01

[a]Adjusted R^2 = .53. Sample size = 110.

The next two factors, age and amount requested, have only a slight relationship to score but are statistically significant. For example, on the average, reviewers gave applicants who had 20 years of experience a worse (higher) score of .2 on a 1-to-5 scale than applicants who had 10. Similarly, applicants who asked for $2 million instead of $1 million were scored lower by one tenth of a point. Lastly, those asking for a longer grant actually tended to get a better rating. Those asking for an additional year had a better score of .36 on the average.

Overall, our model explained more of the variance in score for NSF (53 percent) than it did for NIH (24 percent). This could mean that factors outside our model, such as the importance of the question and the strength of the design to answer that question, were somewhat less important at NSF than NIH. Of course, race data were available only at NSF, so the lower variance explained at NIH may stem in part from not having race in the equation. We return to this issue below.

Factors Related to Scores at NEH

At NEH, several factors were related to scores, as shown in table 3.8. Most important were the series of dummy variables measuring a reviewer's perception of the applicant's track record. Once again, the stronger the perceived track record, the better the score. These results confirm the apparently consistent pattern of potential Matthew effect across all three agencies: all things being equal, proposals from those with stronger track records tend to fare better in peer review.

Table 3.8: Factors Related to Proposal Scores at NEH[a]

Independent variable	Regression coefficient	Standard error
Constant	3.27	.41
Applicant is perceived as major figure	-1.21	.43
Applicant is perceived as notable contributor	-.89	.36
Applicant is perceived as merely competent researcher	-.69	.36
Reviewer is directly familiar with applicant	-.59	.38
Reviewer is indirectly familiar with applicant	-.46	.23
Applicant is male	-.72	.29
Requested amount (per $1,000)	.04	.02

[a]Adjusted R^2 = .35; sample size = 61.

A second set of factors at NEH concerned whether a reviewer had either direct or indirect knowledge of the applicant. Direct knowledge was associated with scores .59 points lower (better), indirect knowledge .46 lower. Again, only those unknown to the reviewer were disadvantaged.

Third, applicants who were men were given on the average much better ratings than women, with scores that were on the average .72 points less

than women's, all other things being equal. Finally, the amount requested was statistically significant but with a small coefficient.

Overall, our model for NEH explains less of the variance in score (35 percent) than that for NSF (53 percent) but more than for NIH (24 percent). One possible interpretation of this finding is that reviewers at NEH give less weight than those at NIH to the key factors external to the model—namely, the importance of the proposal question and the strength of the design to address that question.

These differences are in line with our observations of peer review panels at the three agencies. At NEH, the panels were smaller and generally included members from several different disciplines, and criteria were less consistently applied than at the other agencies. In general, initial scores on individual proposals tended to be quite divergent, often spanning the entire range of possible scores. We also noted that about one third of the preliminary reviewer ratings changed after panel discussions, often moving by several categories. In contrast, at NIH even small changes, on the order of a few tenths of a point, were made grudgingly. We believe these findings and observations mean that peer review at NEH, with its small panels and broad fields, is less robust than at NIH. As one critic has observed, peer review works better within than across fields: "the broader the intellectual territory covered, the less consensus there will be on ranking."[24] In fact, on several occasions at the panels we attended, panelists were sufficiently puzzled by a proposal to ask the program officer and one another, "Is this humanities?" One member of our advisory group told us that history and literature are at present very divided disciplines in which criteria for judging the importance of a question are disputed, or even rejected as useless, and the concept of research design hardly exists. The picture that emerges is one of relatively uncertain reviewers applying whatever information they can to inform their evaluations.

Reviewers' Comments From Our Survey

When asked what in their view were the benefits of the peer review process, reviewers generally commented that the system at their agency was fair or at least the best available. Several likened it to democracy, saying that it was "the worst possible system except for all the alternatives." NIH reviewers were the most likely to say their procedures were fair (42 percent, versus 16 percent at NSF and 32 percent at NEH). NSF

[24]Harvey Brooks, "The Problems of Research Priorities," Daedalus, 107:2 (Spring 1978), 178.

reviewers were more likely to mention that proposals get expert review (23 percent) than that peer review was fair.

However, when asked what in their view were the drawbacks of peer review at the agency, the great majority of reviewers noted at least one drawback, or problem, in peer review procedures. In fact, 98 percent of NIH reviewers reported at least one problem, as did 92 percent at NSF and 77 percent at NEH. However, virtually no one suggested replacing peer review. The criticisms were in the spirit of reforming an essentially viable, if imperfect, process.

Conclusions

Overall, the evidence using multiple regression models is that much of the variation in peer review scores of proposals at all three agencies cannot be explained by the factors we could measure. Indeed, some factors alleged by critics to affect peer review scores, such as an applicant's age or academic rank, do not show up even in simple bivariate relationships. We have interpreted this to mean that scores are driven by agencies' criteria, including the importance of the issue and the quality of the research design.

However, some data relationships did emerge that may lend some support to allegations of bias. In particular, we found evidence that a reviewer's personal familiarity with an applicant was associated with better scores at both NSF and NEH. Furthermore, gender was related to scores at all three agencies, and race was significantly related to scores at NSF. While it is possible that the proposals of women at all the agencies and of nonwhites at NSF were weaker than those of whites and males at the respective agencies, both of these findings could also support the possibility of bias. There is, however, much less ambiguity about the use of unwritten decision rules and lapses in enforcement of conflict-of-interest rules we witnessed at panel meetings. Finally, an applicant's previous track record was an important factor at all three agencies.

It is important to remember that we did not have data on the quality of the proposed research designs, nor on the importance of the questions to be studied by the proposals, because the agencies use only a single summary rating of the entire proposal. A not too surprising consequence of this was that at no agency did the models we fit explain more than about half of the variance in scores. We have generally interpreted this to mean that the major factors influencing scores were those of the contents of the proposals themselves, principally the questions raised and the designs.

This finding varied across agencies. Reviews at NIH appeared to be least vulnerable to considerations such as the Matthew effect and personal relationships, those at NSF and NEH most vulnerable. We have suggested that this could reflect differences in the makeup of the panels and the breadth of subjects they must cover. However, most of our findings may be explained by rival hypotheses. For instance, it could be that much of this unexplained variance is random error. In other words, it is possible that an important factor determining scores is luck.

Finally, we can conclude that several factors alleged to affect reviews are in fact not related to scores at any of the agencies. Neither a reviewer's proximity of expertise nor his or her knowledge of the literature was related to scores. Nor was an applicant's region or academic rank related to the scores his or her proposal received. Even the prestige of the applicant's department had no relationship to the score at any of the agencies. So while we have to be cautious in our interpretation of what factors are related to scores, our model allows us to clearly dismiss several alleged factors as influences on reviewer scoring.

NIH, NSF, and NEH Award Decisions

In the previous chapters, we looked at the factors related to the selection of reviewers and the scores peer reviewers give. While strictly speaking peer review ends with advice, in the form of written comments and scores, given to administrators on which grant proposals to fund, exactly how those administrators apply that advice is crucial and controversial. It is crucial because peer review is about not just the allocation of proposal ratings but the proper and effective allocation of funds. The awards stage is controversial because of perceptions that program officers have trouble denying their friends or that, even when scores are equivalent, insiders are more likely to be awarded grants.[1] The best peer review process could be subverted by program officers and administrators, or, conversely, problematic peer reviews could be compensated for by conscientious administrators.

So our last major question is, "What factors do program officers and administrators consider in their funding decisions?" Since the unit of analysis now shifts from the individual reviewers to the proposals, for each application in our sample we combined administrative and survey data relating to all reviewers of that application to provide average measures for each variable. For example, for reviewer scores at NSF and NEH, we used the mean of the individual reviewer scores given to each proposal; at NIH, we were able to use the overall percentile score, which combines scores from all panelists.

Funding Award Policies and Procedures

The policies and procedures for awarding research grants vary by agency. At NIH, the scientific review administrators do not make funding decisions. Rather, the initial review group's written recommendations, priority and percentile scores, and funding recommendations are forwarded to the institutes for final funding decisions. Each of the 20 institutes and centers providing research support has a national advisory council that, among other things, formally recommends to the directors the proposals that the institutes should fund. However, the councils generally have hundreds of proposals before them and in fact only scrutinize about 20, usually proposals brought to them by institute staff. So the principal role of advisory councils is not conducting peer review of proposals but, rather, setting a recommended percentile cutoff point for proposals to be funded. While the institutes can compensate for the limitations of priority scores

[1]George T. Mazuzan, "Good Science Gets Funded," Knowledge, September 1992, p. 70. Adil Shamoo found that the NIH advisory council members' scores were about the same as nonmembers', yet council members enjoyed a higher award rate. He speculates that this may stem from council members having insider knowledge of which institutes are least competitive and fund the highest proportion, or percentile, of applications.

by considering the summary statements, they generally do not. As the division of research grant's director, Jerome Green, recently told the journal Science, "Institutes tend to follow too slavishly to percentile."[2]

In contrast, NSF program officers are given considerable discretion in funding decisions. After the proposals are reviewed, the program officers evaluate them in light of the reviewers' comments and ratings. They may also consider other factors, such as whether a proposal addresses underfunded areas of research or comes from an applicant at an institution that has not been heavily supported in the past or concerns innovative, high-risk research. The program officer's recommendations are then reviewed at one or possibly two higher levels (section or division) before a final decision is made. After all references identifying the reviewers are removed, verbatim copies of the reviews are automatically sent to principal investigators. Further information can be obtained upon request. However, unlike at NIH, these typically are not provided before final action is taken, so applicants who believe their proposals were not fairly considered must resort to a post hoc appeals process or wait for the next funding cycle.

There are many similarities between NSF and NEH. Like NSF, the program officer at NEH has a fair amount of discretion and uses the panel recommendations and ratings to select a list of proposals to fund. In fact, the program officers at NEH may have more leeway than those at NSF because peer review panels are discouraged from considering the proposal's budget and costs in their evaluations. The program officer's recommendations are then reviewed internally at the next level of management and forwarded to the National Council on the Humanities and its relevant subcommittees. After the council makes its recommendations, the chairperson of NEH makes the final award decisions.

Unlike either of the other agencies, the programs we evaluated at NEH do not routinely provide feedback to applicants.[3] Although such feedback is available on request, officials of several NEH programs told us they do not encourage applicants to ask for feedback. In part, this may reflect budget constraints and the costs of providing feedback on small grant requests.

[2]"Conflicting Agendas shape NIH," Science, 261 (September 24, 1993), 1679.

[3]Some of the program divisions such as the division of public programs do routinely reply to rejected applicants.

Factors Related to Award Decisions

If peer review is to be meaningful, it should generally be the most important factor in an award decision. Other factors might be considered by decisionmakers, such as recognizing unusual problems with the review of a particular proposal, filling a gap in the program's portfolio of grants to meet the agency's mission, or correcting some other imbalances among the awards. Award decisions might also consider track record and institutional support. However, since some of these factors have already been weighed into the ratings by peer reviewers, any further consideration of them means they have been counted twice and their overall effect on the proposal's fate has been greatly amplified.

As with our analysis of ratings factors, we begin our analysis of award factors by looking at some of the bivariate relationships between award decisions and factors alleged by critics to affect them unduly. Then we report on statistical models of all these factors taken together. As with the bivariate tables on factors related to score in the previous chapter, the bivariate tables for factors related to awards will provide some solace to critics. There are reasons one might perceive bias when considering factors in isolation. However, scores and awards are based on complex and interrelated criteria. So the bias that seems apparent when looking at factors one at a time can be quite misleading.

The tables that follow report award percentages for different categories of applicants. Since our survey selected an approximately equal number of proposals that were funded and that were not, the award rate for the sample was also about 50 percent.[4]

The Halo and Matthew Effects and the Likelihood of Being Funded

A common criticism of peer review is that applicants from prestigious schools are given undue preference because of the halo effect. Table 4.1 shows how applicants' perceived departmental prestige is related to actual success in getting funded. What we use here is our reviewers' perceptions of departmental prestige rather than the departmental rankings by NAS. (What we care about are current perceptions in the field. Since we do not have a measure of program officers' perceptions, we use the perceptions of their peers, the panel reviewers, as an approximation.)

[4]When we categorize applicants by another variable, as we do in the following tables, the exact percentage of awards and declinations for all relevant cases varies because of missing responses on the independent variable, and the range is about 45 to 55 percent of proposals actually funded.

Table 4.1: Applicants' Department Rank and Percent of Applications Funded[a]

Agency	Among top 5	Among 6th to 20th	Not among top 20
NIH	60%	53%	47%
NSF	78	44	38
NEH	67	36	42

[a]Sample sizes are NIH = 82, NSF = 92, NEH = 52.

The relationship between funding success and departmental prestige is weakest at NIH, where applicants perceived to be from the top 5 schools had a success rate 13 percentage points higher than applicants from unranked departments. NEH had larger disparities in success rate, but NSF had by far the largest, with proposals from the top 5 schools enjoying about twice the success rate as those from the other categories. Note that this effect was concentrated among applications from departments ranked by reviewers as among the top 5; they did distinctly better at all the agencies. In general, the second tier, those from institutions ranked from 6th to 20th, had success rates that were more like those of the less prestigious departments than like the top 5.

Another point of controversy is the Matthew effect (discussed in chapter 3)—that is, the influence of track record. Table 4.2 shows that track record was related to funding decisions at all three agencies. However, track record was less closely related to success at NIH and NSF than at NEH. The relationships across all three agencies were actually quite similar, except that unknown researchers had a much lower success rate at NEH. Note that applicants who were perceived as "competent" were less likely to be successful at NSF and NIH than applicants whose work was largely unknown. Reviewers and program officers in those agencies may be more willing to take a chance on an unknown, but potentially highly productive, scholar than on someone perceived as merely competent from whom they expect only modest results.

Table 4.2: Applicants' Track Record and Percent of Applications Funded[a]

Agency	Major figure or notable contributor	Competent researcher	Unknown researcher
NIH	62%	34%	40%
NSF	63	35	43
NEH	65	37	17

[a]Sample sizes are NIH = 82, NSF = 92, NEH = 52.

Personal Familiarity and the Likelihood of Being Funded

We did not have a direct measure of the personal familiarity of program officers and other funding administrators with applicants. However, an applicant who is well known to reviewers may also be better known to program officers. So we used the evidence of a reviewer's average familiarity with applicants as a general measure of how well known an applicant was.

We found that being well-known was related to funding at NSF and NEH but not at NIH, where well-known and unknown applicants were equally likely to be funded. As shown in table 4.3, at NSF the best known applicants were three times as likely to be funded as unknown applicants and succeeded twice as often as the moderately well known. A similar, though weaker pattern, appeared for NEH.

Table 4.3: Applicants' Being Well-Known and Percent of Applications Funded[a]

Agency	Well-known applicant	Moderately well-known applicant	Unknown applicant
NIH	50%	56%	49%
NSF	68	37	23
NEH	57	46	37

[a]Sample sizes are NIH = 88, NSF = 92, NEH = 53.

Gender and the Likelihood of Being Funded

Proposals from women were less likely than those from men to be funded at NSF and NEH and more likely to be funded at NIH (table 4.4). The funding rate gap was greatest at NSF, reaching 30 percentage points, and proposals from men were more than twice as likely to be funded as those from women.

Table 4.4: Applicants' Gender and Percent of Applications Funded[a]

Agency	Female	Male
NIH	73%	46%
NSF	22	52
NEH	33	59

[a]Sample sizes are NIH = 99, NSF = 91, NEH = 82.

Conclusions About Bivariate Relationships in Funding

The evidence above of bivariate relationships shows why some observers have inferred that influences we tested exist. However, bivariate observations can be quite misleading because they can be spurious and are

likely to be related to one another. For instance, top departments probably get to be that way by recruiting scholars with strong track records and by providing an environment conducive to research productivity, including superior research facilities, modest teaching loads, and graduate students with research skills. Consequently, it would be surprising if the applicant's departmental prestige and track record were not interrelated. Similarly, well-published scholars may also be expected to present more papers, participate in more conferences, and be asked to do more public speaking than their colleagues. So a strong track record probably results in a scholar's becoming more widely known throughout the profession. Sorting out such relationships and their effects on scoring can be complicated. Thus, in the next section we examine the results drawn from a set of statistical models designed to control for some of these complications.

Logistic Regression Model of Funding Award Decisions

As with our analysis of reviewer scoring of proposals, we developed statistical models of factors that could affect the final decision to fund individual proposals. The models we developed were based on logistic regression, a statistical technique suited to modeling of dichotomous dependent variables—that is, variables with only two values, such as the decision to fund or not to fund a proposal. We tested the variables discussed in the last section along with other possible factors, such as an applicant's region and the amount requested in the proposal. The regression was calculated by entering all the variables at once and sequentially eliminating the nonsignificant ones from the model.

Table 4.5 summarizes the models for all three agencies. At NIH, the only independent variable having any statistically significant power for predicting a grant award was the percentile score given by the peer review panels of the initial review groups. This, however, was a very powerful predictor and confirms what we observed from our site visits—that awards are based on the percentile guidelines set by the advisory councils and that the institute administrators making the final decisions adhere very closely to these guidelines.

Table 4.5: Logistic Regression Model of Funding Decisions[a]

Agency	Independent variable	Beta	Standard error	Chi square[b]
NIH	Constant	15.87	7.12	
	Score	-5.30	2.27	96.5
NSF	Constant	5.86	1.37	
	Score	-1.93	0.49	
	Amount x familiarity	-0.06	0.03	37.2
NEH	Amount requested	2.02	0.85	
	Amount requested x score	-1.02	0.43	54.0

[a]Sample sizes are NIH = 76, NSF = 71, NEH = 52.

[b]Each model has 2 degrees of freedom. All relationships are significant at $p < .05$.

At NSF, score also was the dominant independent variable affecting the funding decision: the worse (higher) the peer review score, the lower the chances of getting funded. However, in addition we found a small, but statistically significant, interaction effect between the amount requested and how well the applicant was known.

A couple of examples can illustrate how the interaction effect works. According to the model results, if a proposal with a top score (1.0) costing $100,000 were submitted by a well-known researcher, the chances are he or she would get an award about 96 percent of the time; even if he or she were completely unknown, there still would be a strong (80 percent) chance of funding. If the requested amount were $500,000, the chances would not change appreciably for the well-known applicant, while the unknown's chances would drop somewhat, to 76 percent. In contrast, if at a requested amount of $100,000, the score drops to 2.0 rather than 1.0, the chances of funding for the well-known applicant would drop from 96 percent to 80 percent, but for the unknown researcher the chances would drop much more sharply, from 80 percent to 37 percent. Thus, other things being equal, the higher the requested amount, the more likely the well-known researcher would be to succeed, and this difference would become wider as the scores got worse.

This result can be interpreted as showing prudence on the part of NSF program officers. That is, in cases in which reviewers gave strong ratings to proposals, little discretion was used to make funding decisions. However, for weaker scores, officers took account of how much was requested; and the more requested, the more unwilling they were to approve a grant to a researcher who was not well known.

At NEH, the amount requested played an important role, with both a direct effect on the chances of being funded and an additional effect through an interaction with score. To illustrate, for a mean score across reviewers of about 1.96 (on a 5-point scale), the odds of getting funded would be even (50 percent), no matter what the amount requested. For more unfavorably rated proposals, the chances of receiving funding would get worse and the chances decline more quickly with larger requested amounts. Thus, for a proposal with a mean score of 3.0, the chances of getting funded would be 35 percent if the requested amount were $10,000 but only 0.5 percent if the amount were $50,000. These results may reflect the tight budget constraints at NEH.

Conclusions

Overall, we found that the most important factor affecting whether a proposal was funded was the score assigned by reviewers, as expected. However, at NSF we also found that the amount requested and being personally well known were, together, related to awards. At NEH, awards were related to the amount requested.

In addition, it is important to note that the overwhelming importance of scores on funding decisions means that any factors affecting scores also affect funding. Thus, our models indicate that the program officers and agency officials responsible for funding decisions do not consider the applicant's gender, department affiliation, or track record. However, as we noted in chapter 3, we did find some relationship between these variables and reviewers' scores at some agencies, and since awards are in turn highly related to scores, these other factors could have an indirect influence on funding decisions in those cases.

Conclusions, Recommendations, and Agency Comments and Our Response

The intent of this report was to examine, as requested by the Chairman of the Senate Committee on Governmental Affairs, the empirical evidence for problems in the equity, or fairness, of peer review at federal agencies. We focused on peer review at NIH, NSF, and NEH and on three critical stages of the peer review process—the selection of reviewers, the review and rating of proposals by peers, and the final decision to fund or not. This is not to say that other issues are without importance; concerns about the efficiency and efficacy of peer review are certainly in need of study as well. Some of these concerns, such as the poor fit between a researcher's need for risk taking and peer review's tendency to be risk averse, are critical questions that go beyond the scope of this report.

Conclusions

Overall, we found that peer review processes appear to be working reasonably well. It is clear from reviewers' responses to our open-ended questions that they believe peer review is the best available method for allocating research funds. Virtually no one suggested replacing it. Yet a majority reported some problems and suggested reforms. We found empirical evidence of potential problems in some areas but not in others, suggesting that agencies need to take a number of measures to better ensure fairness in the three areas of the study's focus.

Reviewer Selection

Our data did not confirm a number of the critiques identified in the peer review literature. First, we found little evidence to support the notion that peer review panels are staffed disproportionately from among researchers at a handful of elite institutions; indeed, elite institutions tended to be somewhat underrepresented among reviewers compared to applicants. Second, although we did find notable disparities in the number of reviewers across regions, this largely stemmed from differences in the underlying applicant base rather than selection bias.

The agencies had more of a mixed record on gender and age. We found that, while women served as panelists at least in proportion to their representation among sample applicants at all three agencies, they were underrepresented among the external reviewers of some programs at NSF. This is all the more important because of NSF's heavy, and in some disciplines exclusive, reliance on outside reviews.

In addition, at all three agencies young scholars and those with lower academic rank were underrepresented. For example, assistant professors accounted for no more than 4 percent of panelists and had only a slightly

better representation of 8 percent among external reviewers. Recent studies of peer review for academic journals show that young reviewers actually give better reviews. This suggests the need for efforts to recruit junior scholars as proposal reviewers, at least experimentally, to ensure their greater representation. Given that some academic institutions actively discourage junior scholars from participating in peer review for federal programs, these agency outreach efforts might involve encouraging institutions accepting agency support to credit such participation in making promotion and tenure decisions.

Finally, at a more basic level, we found that there was some tension between the ideal of a high level of reviewer expertise on the topic of a research proposal, on the one hand, and a low level of personal familiarity with the applicant, on the other. That is, all things being equal, one might assume that the fairest reviews would be given by those who are both highly knowledgeable in the specific area of inquiry in the proposal and personally and professionally disinterested in the applicants. But, in fact, those who know the subject matter best are most likely to know the applicants, creating the potential for personal and professional considerations to color the reviewers' judgments.

These problems are most clearly seen by comparing NIH and NSF. Reviewers for NIH reported less proximate research expertise and also less personal familiarity with the applicants than reviewers at NSF. In contrast, NSF reviewers indicated more proximate research interests, and they were also more likely to know the applicants.

These differences appear to reflect the traditions of peer review at the two agencies. NIH tends to emphasize a problem area (such as AIDS or aging) and to include reviewers from different disciplines with interest in the area on its panels. By contrast, NSF reviewers are selected largely by discipline (such as theoretical physics or economics). The perhaps unsurprising consequence is that NSF reviewers are more likely than those at NIH both to be working on questions closely related to those of the applications they are reviewing and to know the applicants personally.

Compounding these differences, NIH relies primarily on panels, while NSF uses mail reviewers more extensively. Because panelists have to evaluate a great many proposals over a broad range of topics, they are less likely than mail reviewers to be working on the same questions addressed in the applications they review and are less likely to have personal, camp, or self-interest conflicts. But they are also less likely to be expert in the

precise subject of the application. In contrast, because mail reviewers are selected to match the topics in the proposal rather than for their general breadth of knowledge, they are more likely to have highly relevant expertise but also to know applicants personally. Panel reviews are also more public than mail reviews, providing at least a partial check on possible biases.

Efforts to alleviate the relatively lower level of immediately related expertise at NIH might lead to an increase in the number of panels or a more extensive use of subpanels, as is now done for very large panels. Alternatively, NIH could more fully integrate the work of ad hoc mail reviewers into that of the panels, for example by disseminating their reviews to all members prior to panel meetings. Conversely, offsetting the potential influence of personal relationships at NSF may require movement in the opposite direction—toward a stronger role for panels.

NEH reviewers fell between those at the two other agencies on measures of relevant expertise while also reporting the least direct personal knowledge of applicants. Given that the small panels at NEH are required to review applications covering wide areas of the humanities, more extensive use of mail reviews might help alleviate the expertise problem without unduly exacerbating the potential for bias arising from personal knowledge of the applicants. In addition, mail reviews would cost less than increasing the size of panels to improve expertise.

Problems in Peer Review Rating Procedures

In examining the factors related to reviewers' scoring of proposals, we found some apparent relationships between scores and other factors that we tested further, using multivariate regression analysis. In this analysis, many of these relationships did not hold, but several did.

We found that ratings were related to gender at all the agencies and to race at NSF. While it could be that these differences stemmed from lower-quality proposals being submitted by women at each agency and nonwhites at NSF, rather than partiality toward men and whites, our data do not allow us to conclude one way or the other. Nonetheless, these findings cannot be used to disprove allegations of partiality, which need to be further investigated and addressed by the agencies. One way to accomplish this would be to compare results of reviews performed in the conventional peer review manner to "blind" reviews for a sample of proposals.[1]

[1]Blind proposals would have all references to name, race, gender, and present or past institutional affiliations removed. Proposals could be sent to some reviewers without such information and to others with the information included, as usual, and ratings compared.

We could not study the relationships between an applicant's race and scores at NIH and NEH because these agencies do not retain data on the race of applicants. This lack of data weakens efforts to ensure nondiscrimination in reviews at these agencies. In addition, NIH data files lack information on scores given by individual panelists. Yet, knowing the actual scores panelists give is virtually the sine qua non of effective evaluation and oversight of peer review. These data could be maintained in a way that protects the information from those with no need to know but that allows full and unrestricted access to these data for congressional oversight. NSF currently follows this practice without compromising the privacy of participants.

We also found at all the agencies that an applicant's track record was related to scores. While an applicant's track record is in one form or another a criterion at each of these agencies, guidelines are uniformly vague about its importance—that is, how much it should be considered in an overall score.

Lack of clarity about what criteria are used and how they are used makes it difficult for all participants to have a precise understanding of how peer review is being applied.[2] This lack of clarity occurs in several ways. First, the agencies do little to ensure that reviewers have an accurate and similar understanding of an agency's criteria and rating scales. The problem is especially severe at NSF, where there are a variety of panel tenure systems, and NEH, where panels convene for one time only, and therefore both agencies lack the continuity of membership that can lead to a consistent application of agency policies. This suggests the need for better calibrating ratings among reviewers.

Second, reviewers are not necessarily considering the agencies' formal criteria consistently. This probably contributes to much of the dissension about and mistrust of peer review. Providing ratings for each criterion, along with a separate overall rating, as is already being done in NIH contract reviews, could help address this problem by ensuring that all criteria are addressed as part of the review. At a minimum, it would provide useful support for feedback to applicants on the areas of relative strength and weakness in their proposals, which could improve their subsequent submissions. In addition, it could stimulate discussions among

[2]As Stephen Cole noted in his discussion of scientific networks and particularism, "perhaps most important, is the clarity of the criteria of evaluation. The more ambiguity in the criteria and the less consensus on them the greater the chance for the operation of all types of particularism, including that of network ties." Stephen Cole, Making Science: Between Nature and Society (Cambridge, Mass.: Harvard University Press, 1992), p. 185.

reviewers that would clarify which criteria are most important for a given proposal or area of research. Recalling that peer review is being called upon to make ever finer distinctions between outstanding and merely very good proposals, rating on individual criteria also may result in more precise reviews and award decisions, and this precision would contribute to a more efficacious use of federal funds. Furthermore, if the agencies want to monitor and improve the quality of research they are funding with taxpayers' money, they will need these data to know which of their criteria are relevant to funding successful science and humanities projects.

Third, we found that unwritten or informal criteria were used by panels at all three agencies. For instance, many of the panels we observed had informal criteria about how much of the preliminary work should already be done at the time of application. This is potentially unfair to new applicants, who lack insider knowledge on such issues. We found no evidence of efforts either to formalize such decision rules or to communicate them to applicants through proposal-writing guidance or other outreach materials.

Factors Related to Funding Decisions

We also considered the factors related to agency decisions to award funding. We found that the most important factor affecting whether a proposal was funded was the score assigned by the reviewers. Score was the only statistically significant criterion related to funding at NIH, but the amount requested was also a factor at both NSF and NEH, and an applicant's being well-known or not at NSF.

Recommendations

From these conclusions, we recommend that the Director of NIH, the Director of NSF, and the Chairman of NEH take the following actions.

First, address the underrepresentation of young, junior scholars as reviewers by targeted outreach efforts, at least experimentally.

Second, increase the monitoring of discrimination, including conducting tests comparing blind to conventional reviews to ensure that gender, race, and ethnic discrimination are not affecting scores given by peer reviewers.

Third, address the lack of clarity in the application of review criteria by developing rating systems in which proposals are rated separately on a number of criteria (such as importance of the issue, quality of the design,

institutional support, and qualifications of the applicant), along with the overall rating.

Fourth, identify any commonly used unwritten or informal decision rules that are applied by peer reviewers and, where feasible, formalize them or at least inform applicants of their importance.

The Director of NIH should take the following additional actions.

First, produce a better match of panelists' area of expertise to that of the proposals by making greater use of subpanels and more fully integrating the work of ad hoc mail reviewers into the panel process.

Second, improve evaluation and oversight by retaining data on scores given by individual panelists and the race and gender of individual applicants.

The Director of NSF should take the following additional actions.

First, address the potential for bias arising out of extensive personal relationships between reviewers and applicants by increasing the use of panels, where feasible.

Second, address reviewer representativeness by more closely monitoring the inclusion of women and minorities among external reviewers.

Third, increase efforts to calibrate proposal ratings among reviewers through information provided in advance and discussions of examples at the convening of panels.

The Chairman of NEH should take the following additional actions.

First, improve the level of relevant expertise in reviews by making greater use of mail reviewers.

Second, improve evaluation and oversight by retaining data on the race of applicants.

Third, increase efforts to calibrate proposal ratings among reviewers through information provided in advance and discussions of examples at the convening of panels.

Agency Comments and Our Response

A draft of this report was reviewed by officials of the Public Health Service (for NIH), NSF, and NEH. Their written comments are reproduced in appendixes II, III, and IV, along with our detailed responses. Here, we summarize the comments related to our recommendations.

Discrimination Issues

Regarding possible discrimination issues, the agencies generally concurred in the importance of protecting against discrimination, but all three cited specific objections to collecting and retaining needed data. NSF agreed to monitor for race and gender discrimination among reviewers but argued that "legal considerations regarding privacy do not allow us to ask about gender or ethnicity of reviewers, or to keep that information in our databases." However, we are aware of no such prohibitions, although reviewers may decline to provide such information. In fact, the data NSF supplied to us included information on the gender of reviewers.

NEH disagreed with the recommendation to collect data on the race of applicants on the grounds that we had not found evidence of discrimination at NEH. This disagreement, however, is specious: the basis of the recommendation is not that we found discrimination but, rather, that there were no data to evaluate whether there was discrimination or not. Unless the data are made available, neither the agency nor the Congress can know whether such discrimination is occurring.

Finally, PHS argued against retaining data on reviewers and scores, except in relation to certain reviews it agreed to conduct (see the next section). This could be an acceptable solution, if the sample of cases is carefully drawn and the study is done at regular intervals, such as annually.

PHS did agree to consider using blind reviews to test for discrimination. NEH said that it, too, was considering the use of blind reviews in one of its programs but did not see a broad utility for such procedures, and NSF rejected the idea outright. However, both NSF and NEH misinterpreted our statements as recommending that blind reviews be used in place of normal reviews rather than in comparison to them, as we intended.

All three agencies concurred in the need for ensuring representation of junior scholars among reviewers, although both PHS and NEH took issue with our characterization of the extent to which they currently use them.

Conduct of Reviews

All three agencies generally agreed with the need to identify any unwritten criteria reviewers may be using. However, PHS disagreed that the case of preliminary results exemplified an unwritten rule at NIH because applicants are encouraged to include such results in their submissions. We disagree with PHS; while NIH instructions describe a discussion of preliminary results as "optional," reviewers treat it an essential requirement. NEH also expressed concern about overly formalizing the review process in attempting to identify unwritten rules. We agree that there are practical limits on agencies' ability to formalize all rules but, in fact, we specify only "commonly used" unwritten rules. Moreover, our recommendation allows the agency to simply notify applicants of these unwritten criteria in lieu of formalizing them.

NSF and NEH also both agreed to improve calibration. NEH argued that it already makes such efforts, but we did not find this in evidence during our observations of NEH panels.

However, all three agencies disagreed with scoring proposals along individual criteria. PHS noted that an earlier NIH study had looked at weighting individual criteria and found that reviewers were "not comfortable" with the idea, although PHS did not indicate the source of discomfort. NSF expressed concern that such a change could lead to "meaningless arithmetical distinctions" and less reliance on reviewer comments. Finally, NEH noted that it had already experimented with such a scoring system for first-stage reviews of 1993 dissertation grants program applications, and while it found this "useful and efficacious," it did not think it would be applicable to other programs (although NEH did not indicate why not). However, the comments clearly demonstrate that the agencies misinterpreted our statements as recommending weighted numerical scoring by criterion. This is not the case. What we are recommending is a procedure that ensures that all elements of a proposal are at least considered by reviewers in arriving at an overall appraisal. We have reworded the recommendation to make this clear.

Finally, PHS objected to reducing the scope of NIH panels, taking issue with the finding of limited relevant expertise among its reviewers. We stand by the finding but have clarified the recommendation. In addition, after considering reviewers' comments, we have added recommendations that NSF address the possible influences of personal relationships between reviewers and applicants by making more extensive use of panels, where feasible, and that NEH make greater use of mail reviewers. These

recommendations were not included in the draft reviewed by NSF and NEH officials.

The agencies had a number of technical comments, which we have addressed where appropriate throughout the report.

Framework of Potential Weaknesses and Agency Actions and Policies

Table I.1: Reviewer Selection: Reviewers' Lack of Relevant Expertise

Potential source of weakness	Agency action and policy		
	NIH	NSF	NEH
Reviewers lack relevant expertise	Must demonstrate through Ph.D. or M.D., publications, honors, seniority, and active research	Guidelines stress pertinent expertise or demonstrated ability	Detailed questionnaires used to ascertain areas of expertise
	Large panels of 16-20 members are designed to cover a range of fields	In some programs, panelists are selected ad hoc after bulk of applications are in	Each panel is selected anew to fit proposal topic
	Applications not fitting standing review group are referred to special ad hoc group	Reviewers should have special knowledge of science and engineering subfields involved	Data on fields of expertise are maintained on computer data base
	External mail reviewers are used only to complement panel's expertise on one or two proposals	In some programs, outside reviewers are selected ad hoc after bulk of applications are in	Research division uses prepanel external reviewers for technical support; other divisions add consultants as needed
	Consultants are added if there are several proposals for which the panel needs special technical expertise	Applicants may suggest reviewers and disapprove reviewers' suggestions	Applicants may suggest up to half of reviewers and can disapprove reviewers' suggestions
	For 100 study sections, reviewers are drawn from 2,300 initial review group members	For 140 research programs in 27 divisions, 60,000 reviewers are used annually from group of 150,000	For 200 panels, 1,000 panelists are drawn from computerized list of 13,000 scholars

Table I.2: Reviewer Selection: Reviewers' Conflict of Interest

Potential source of weakness	Agency action and policy		
	NIH	NSF	NEH
Reviewers have conflict of interest	Panelists are required to read conflict-of-interest letter, certify they will observe confidentiality rules, list any conflicts, and sign	Conflict-of-interest rules are published (Manual 15)	Reviewers are sent a copy of conflict-of-interest statement
	Reviewers are asked not to participate in cases involving personal conflicts	Family members, close friends, open antagonists, recent advisers, and collaborators are prohibited from reviewing proposals	Reviews are prohibited by those with interest, such as applicant's adviser, principal investigator, or employer
	Reviewers' own proposals are sent to another initial review group	Reviewers cannot have pending NSF proposals in same area or a recent declination	
	Staff member tracks known conflicts and notifies chair before proposal is discussed	Working in the same area of research is not a conflict of interest; reviewers are asked to report if they are uncomfortable and feel they cannot be objective	
	Consultants to initial review group must certify no conflict of interest	Program officers must reveal conflict of interest	
	Panelist from same institution or having a collaborative relationship with applicant must leave the room	Panelist must leave the room if from same institution (enforced by program officer and panelists)	Panelist with a conflict must leave the room if from same institution (enforced by program officer and panelists)

Table I.3: Reviewer Selection: Reviewers Not Demographically Representative of Their Field

Potential source of weakness	Agency action and policy		
	NIH	NSF	NEH
Reviewers are not demographically representative of their field	Rotates one fourth of membership of standing panels and council each year	Rotation varies from program to program	Starts each panel anew; no more than 20 percent of panelists should have served for over 3 years
		Members should be drawn from as broad a set of geographical areas as feasible	Panels should be geographically diverse
	Selection criteria specify "adequate" representation of women	Considers race, gender, and age after expertise established; special attention should be paid to attaining qualified persons from underrepresented groups such as women and minorities to serve on panels	
	Selection criteria specify "adequate" representation of minorities	Members should be selected from as broad a range of age groups as feasible	Panels must reflect broad cultural diversity
		Guidelines call for representation of small, medium, and large institutions, as well as nonacademics and avoidance of concurrent or successive appointments from the same institutions	
		Routinely analyzes composition of panels but not mail reviewers	

Table I.4: Peer Review: Halo and Matthew Effects

Potential source of weakness	Agency action and policy		
	NIH	NSF	NEH
Halo effect: applicants are given better rating because from elite universities	Explicitly assesses institutional resources: one of the principal criteria is the reasonable availability of resources necessary to do the proposed research	Evaluation criteria include explicit assessment of the adequacy of institutional resources	
		Asks reviewers to recognize circumstances at nonelite schools	
Matthew effect: applicants' reputation is given undue consideration	Explicitly evaluates applicants' training and track record in preliminary reviews	Considers past performance in evaluating capability of applicants	Criteria vary but typically consider past performance in evaluating applicants' promise
			Program officer corrects panelists who substitute reputation for merit

Table I.5: Peer Review: Information Used at Panel Meeting

Potential source of weakness	Agency action and policy		
	NIH	NSF	NEH
Incorrect, derogatory, or defamatory information is discussed at panel meeting		Program officers are directed not to use such information	Panel chair or NEH staff direct panelists to disregard inappropriate comments and put that directive in written notes available for review by applicant
	Decision can be deferred if more information is needed	Inspector General is notified and principal investigator notified if inspector approves	Principal investigator can request written comments after decisions
	Principal investigator can request corrective action if error found	Principal investigator can respond and ask for reconsideration	

Table I.6: Peer Review: Reviews Affected by Demographic or Regional Biases

Potential source of weakness	Agency action and policy		
	NIH	NSF	NEH
Reviews are affected by demographic or regional biases		Review criteria include effects on national education and human resource base	
	Funds outreach programs for women, minorities, and persons with disabilities	Funds outreach for applications, particularly from women and minorities	Has small outreach program
	Publishes Guide for Grants and Contracts; also has publicly available data base	Sends newsletter to 20,000 institutions	

Table I.7: Funding Decision: Excess Program Officer and Administrative Discretion

Potential source of weakness	Agency action and policy		
	NIH	NSF	NEH
Program officer and administrative discretion is excessive	Reviewers' preliminary write-ups include award recommendation	Program officer is responsible for final recommendation for each proposal	Program officer makes funding recommendation
	Funding decision is made by separate program staff	Every award is reviewed by division director	Council reviews funding recommendations; chair makes all final decisions
	Council reviews and approves institute funding decisions	Large awards approved by higher authority	Council reviews all proposal actions
		Conducts quarterly reviews and small sample compliance reviews; comprehensive review of each program made every 3 years by visiting committee is available to the public	Council may change panel recommendation if evidence of bias; director has to justify any changes from council action

Table I.8: Funding Decision: Halo and Matthew Effects

Potential source of weakness	Agency action and policy		
	NIH	NSF	NEH
Halo effect: applicants from prestigious universities are given preferential treatment	Has program for baccalaureate institutions not historically major participants in NIH programs	Gives Research Opportunity Award as an add-on at discretion of program officer to help researchers from small institutions	
		Conducts Undergraduate Institutions Program	
Matthew effect: past success is rewarded resulting in a cumulative advantage in competition for future grants	Offers Institute Training Grant for pre- and postdoctoral research training		Guidelines for some programs specify preference to applicants not awarded major grants for previous 3 years
	Offers National Research Service Post-Doctoral Fellowships		

Table I.9: Funding Decision: Demographic and Regional Biases

Potential source of weakness	Agency action and policy		
	NIH	NSF	NEH
Demographic and regional biases are in effect	Offers awards for newly independent researchers	Offers Research Initiation Awards for new Ph.D.s: Presidential Young Investigator Awards	Guidelines stress cultural and geographic diversity in awards
		Internal funding targets set by budget office	
	Conducts several minority-related programs, such as Minority Access to Research Careers	Conducts Minority Research Initiative, Research Improvement in Minority Institutions, Minority Research Centers of Excellence, Research Centers for Minority Scholars, Alliances for Minority Participation	Conducts Faculty Graduate Program for Historically Black Colleges and Universities
	Offers Minority Biomedical Support Grants	Offers Research Opportunities for Women Program, Visiting Professorships for Women, Faculty Awards for Women	
	Reports annually on distribution of awards by state	Conducts program studies of demographic distribution of awards	
	Race and gender are indicated on separate form	Race and gender are indicated on separate form	Race and gender information are not gathered except when requested by the Congress

Table I.10: Efficiency of Peer Review: Inadequate Feedback and Appeals

Potential source of weakness	Agency action and policy		
	NIH	NSF	NEH
Feedback and appeals processes are inadequate	Principal investigator is sent copy of summary statement with percentile rank, priority score, and narrative evaluation	Verbatim reviews are sent without reviewer name and affiliation; program officer explanation sent automatically	Detailed "why not" letters sent by program officer, including suggested improvements that applicant can adopt and resubmit next time
			Verbatim reviews without reviewer name and affiliation sent only upon applicant's request
	Principal investigator and institution are notified within 30 days of council action	Plans to cut review time to 6 months	
	Principal investigator can rebut and appeal serious errors	Principal investigator can request reconsideration of a program officer's decision; a second request can be made to deputy director	Applicant can talk to any staff member, but since decision is made by chair of NEH, resubmission is most effective

Table I.11: Efficiency of Peer Review: Barriers to Application

Potential source of weakness	Agency action and policy		
	NIH	NSF	NEH
Barriers to application exist	Research plan limited to 20 pages and appendixes restricted	Simplified, shortened, and standardized applications required; suggested limit of 15 pages	Pages limited in each program
	NIH Guide for Grants is on Bitnet	E-mail templates are used; E-mail panels are used experimentally	
	AIDs proposals are processed in 6 months	Plans to reduce review time from 9 months to 6	
	Offers outreach publications, videotapes, and presentations by NIH staff	Offers outreach programs	Conducts small outreach program, such as seminars on preparing grant applications in underserved areas
	Contact with NIH staff is encouraged		

Table I.12: Efficacy of Peer Review: May Not Produce Good Science or Humanities

Potential source of weakness	Agency action and policy		
	NIH	NSF	NEH
Peer review may not produce good science or humanities	Past performance is reviewed for renewal and supplemental applications	Conducts occasional program surveys of results	
	Program staff monitor research progress	If applicants have had an NSF grant in the past, reviewers are asked to comment on its quality	
		Is investigating possibility of routine measure of outcomes	

Comments From the Public Health Service

Note: GAO comments supplementing those in the report text appear at the end of this appendix.

DEPARTMENT OF HEALTH & HUMAN SERVICES — Public Health Service

Rockville MD 20857

APR 4 1994

Ms. Eleanor Chelimsky
Assistant Comptroller General
General Accounting Office
Washington, D.C. 20548

Dear Ms. Chelimsky:

Enclosed are the Public Health Service's comments on your draft report, "Peer Review: Fairness in Federal Agency Grant Selection." The comments represent the tentative position of the PHS and are subject to reevaluation when the final version of this report is received.

The PHS appreciates the opportunity to comment on this draft report before its publication.

Sincerely yours,

Anthony L. Itteilag
Deputy Assistant Secretary
for Health Management Operations

Enclosure

COMMENTS OF THE PUBLIC HEALTH SERVICE (PHS) ON THE GENERAL ACCOUNTING OFFICE (GAO) DRAFT REPORT, "PEER REVIEW: FAIRNESS IN FEDERAL AGENCY GRANT SELECTION PROCESSES," FEBRUARY 7, 1994

General Comments

The PHS appreciates the opportunity for review of the GAO draft report and offers the following comments for your consideration.

This GAO report represents a reasonable attempt to address a number of questions often raised regarding the fairness of the peer review process. However, we recommend that a more balanced perspective, particularly in the Executive Summary, be provided to reflect the difficulties in study design which we believe have led in some cases, to inaccuracies and misinterpretation of data in the report. Although the report addresses its limitations in the INTRODUCTION (pages 1-22 & 1-23), the limitations identified are neither the only nor necessarily the major ones.

Now pp. 24 and 25.

See comment 1.

Specifically, the Executive Summary should identify to the reader two other design characteristics of the study: (1) that the National Institutes of Health (NIH) sampling survey was done more than a year after the peer reviews were completed, so that there was heavy reliance on the memory of the reviewers; and (2) sampling was done of NIH review panelists only, without emphasis to the fact that nearly all review panels rely additionally on a varying number of ad hoc mail reviews, solicited precisely because some relevant expertise might be lacking or underrepresented on the panel. The latter design characteristic, for example, leads to the erroneous conclusion that NIH reviewers are significantly less familiar with the science being reviewed than are the reviewers for NEH and NSF.

See comment 2.

In fact, for the past year alone, chartered Division of Research Grants (DRG) study sections had approximately 1,800 ad hoc reviewers present at study section meetings to provide needed expertise in specific areas. In addition, over 4,000 reviewers were used to conduct strictly ad hoc meetings (e.g., for review of members' applications, Small Business Innovative Research Awards (SBIRs), and other special initiatives). Finally, approximately 1,600 members of the community provided written (mail) opinions for consideration by the various study sections at their meetings.

See comment 3.

In addition, the data provided in most of the tables are incomplete in not providing numbers as well as percentages. Consequently, it is impossible to obtain a clear understanding of the validity of the data. In other instances, the data

See comment 4.

Now p. 82.

See comment 5.

Now pp. 79-80.

Now p. 52.

reflect some inaccuracies, such as percentages not totalling to 100% (cf. Table 2.3). In other instances, (e.g., Tables 3.1 and 3.2) no indication of statistical significance is given to the numbers.

Finally, a number of points and/or recommendations appear to reflect a philosophical difference between the NIH peer review policies and the GAO position. For example, on page 5-5 of the report, the point is made that NIH data files lack information on scores given by individual panelists. In fact, NIH procedures are intentionally designed in this manner, because of concerns for confidentiality and reviewers' anonymity.

GAO RECOMMENDATION

We recommend that the Director of NIH, the Director of NSF, and the Chairman of NEH take the following actions:

(1) Address the exclusion of young, junior faculty as reviewers by targeted outreach efforts, at least on an experimental basis.

PHS COMMENT

We concur that the NIH does not involve a large number of junior faculty in the peer review process, but we do not agree that they are excluded. In fact, in many areas of research (e.g., molecular biology and other newly burgeoning areas of research) it is often the younger investigators who are at the forefront. These individuals are frequently used as ad hoc panelists or to provide ad hoc mail reviews. (On this point, for example, the report language on page 5-3, paragraph 1, seems to be based on an imprecise understanding of the NIH process.) In the final analysis, expertise to provide informed scientific and technical merit evaluation is the basis and the defining characteristic of NIH peer review, and the development of such expertise often requires time and maturation. This precise point is alluded to in the GAO report on page 2-49, paragraph 2. At the very least, the word "exclusion" in the GAO recommendation should be replaced with the word "underrepresentation."

(2) Increase the monitoring of discrimination, including conducting tests using "blind" reviews to assure that sex, race and ethnic discrimination are not affecting scores given by peer reviewers.

PHS COMMENT

We concur that this is an important issue and bears monitoring. We propose to design a study to address this concern. The report suggests a possible approach that we will consider: providing some reviewers with unmodified applications and other reviewers with applications from which references to name, race, sex, and present or past institutional affiliations have been removed.

(3) Address the lack of clarity in the application of scoring criteria by considering use of a scoring system in which proposals are rated separately on a number of criteria along with the overall summary score.

PHS COMMENT

See comment 6.

We concur that there is a lack of specificity in the application of scoring criteria, but do not concur that this issue has not been addressed to date, nor do we believe that it needs to be addressed further. The report notes on page 5-6, paragraph 2, that NIH reviewers are not advised about the relative weight that should be given to the agency's formal criteria, except for the review of contract proposals. More accurately, the report notes on page 3-1, paragraph 2, that there is, in fact, no formal weighting system given to grant review criteria. In quoting from an NIH report, the GAO report notes that the relative importance of the criteria may, in fact, vary among applications.

Now pp. 82-83.

Now p. 53.

As a result of the 1988 report of the NIH Peer Review Committee, a study of the weighting of review criteria was undertaken at NIH. The results confirmed that reviewers were not comfortable with the assignment of specific weights to individual review criteria, but felt instead that weighting might vary depending on the application. For example, the assessment of adequacy of methodology might be of less concern when proposed by an investigator who has been consistently successful in the past in implementing similar methodologies to that proposed in the application under consideration. In another application, creativity and innovation might be so prominent as to outweigh several other factors.

(4) Identify any commonly used unwritten or informal decision rules that are applied by peer reviewers, and where feasible formalize them, or at least inform applicants of their importance.

See comment 7.

Now p. 56.

Now p. 84.

See comment 8.

PHS COMMENT

We concur that the use of unwritten or informal decision rules is inappropriate. However, it is not clear that the GAO position on this issue is well founded. It would appear that the only example of an unwritten review criteria used by the NIH which is identified in the GAO report (page 3-7) is the evaluation of preliminary results. However, the case can logically be made that preliminary results are the underpinning for significance, originality, and feasibility of methodology, and a necessary component for their assessment. Further, in the PHS 398 grant application kit, the applicant is specifically instructed on page 21 to provide information on preliminary studies which will help to establish the experience and competence of the investigator to pursue the proposed project. Thus, assessment of preliminary results is a part of, not separate and beyond, existing and published review criteria.

There are two additional recommendations (page 5-9) addressed specifically to NIH.

GAO RECOMMENDATION

The Director of NIH should take the following additional actions:

- First, act to produce a better match of panelists' area of expertise to that of the proposals by reducing the scope of panels or sub-panels, and making more extensive use of outside reviewers.

PHS COMMENT

We do not concur that there is currently an inadequacy in the match of panelists' areas of expertise to the applications under review. As documented above, the use of ad hoc panelists and ad hoc reviews is much more common than acknowledged in the GAO report. Also, the reduction in scope of panels would seem to work against a major positive attribute of the NIH peer review system identified in the GAO report, i.e., the reduced potential for bias in the NIH system. Philosophically, the NIH peer review system has been based upon the use of broader panels than those of NSF and there appears to be no valid reason presented to warrant altering this philosophy.

- Second, improve evaluation and oversight by retaining data on scores given by individual panelists, and the race and gender of individual applicants.

See comment 9.

PHS COMMENT

We concur that the analysis of scores given by individual panelists is useful for purposes of evaluation and oversight. In fact, analyses of these data are performed each review round to guard against scoring improprieties (e.g., "blackballing") and to provide feedback to reviewers. Data concerning individual ratings and group rating behavior are provided. However, we do not concur that retention of data on scores given by individual panelists on individual applications is necessary or desirable, for reasons described above under "General Comments." If the intent is to provide for improved evaluation and oversight related to scores given, for example, to applications submitted by ethnic minorities, women, and researchers from less "prestigious" universities, implementations of Recommendation (2) above, with which we concur, can be designed so as to provide that type of information.

The following are GAO's comments on the April 4, 1994, PHS letter.

GAO Comments

1. The discussion of our survey of NIH reviewers in chapter 1 has been expanded to note this limitation. However, we also point out that a number of NIH respondents we spoke with told us they maintained systematic records on all proposals they reviewed, including their comments and the ratings they gave. This reduces the risk that their responses are based on faulty memories. In addition, respondents who could not recall how they rated the proposals were directed to skip this part of the survey instrument.

2. As we note in the report, we did not have data on mail reviewers at NIH. In fact, we made numerous efforts to obtain these data, but NIH eventually told us it did not maintain them. This may explain why none of the internal NIH assessments of peer review we examined included data on mail reviewers. It is important to point out that mail reviewers do not provide scores on applications, and their written comments are generally not distributed to panel members. As a result, they have little effect on scores, which largely determine funding decisions at NIH.

3. Numbers of cases are now shown on tables and figures where appropriate. We have indicated that all relationships discussed in the text are significant at the .05 level, unless we indicate otherwise.

4. Lack of information on scores given by individual panelists hinders oversight and evaluation of the review process at NIH. In fact, one NIH official reported at a professional conference that "NIH still has almost no analytic data bases; getting one put together is like pulling hen's teeth." We further note that both NSF and NEH maintain such data, apparently without creating confidentiality concerns.

5. We have changed "exclusion" to "underrepresentation." However, it is noteworthy that NIH has a specific policy discouraging the selection of assistant professors as study section members.

6. We have reworded our recommendation to make clear that we do not propose a relative weighting of elements covered in applications. What we are suggesting is a review format that would ensure that all elements of each proposal are considered before an overall evaluation is reached. Establishing that proposals meet some minimum rating for each element could help identify proposals too weak to merit funding. In addition, while

we recognize that different reviewers will place different emphases on each element, separate scoring should inform the discussion among panel members and provide useful feedback to applicants, which might improve future proposal submissions. The latter consequence would be especially important for researchers new to the NIH funding system.

7. The need for preliminary results was the main unwritten rule we observed being used by NIH reviewers. While the NIH guidelines for reviewers specifically call for evaluating preliminary results, the directions to applicants mark these as "optional" for inclusion in the proposal. Our observations at NIH panel meetings and discussions with NIH officials indicate that panelists look carefully at preliminary results, so that applicants who do not fully address this issue apparently place themselves at a disadvantage. In fact, a privately published guidebook on writing grant applications has this advice: "Although this section is 'optional' according to the NIH instructions, it is very important that you provide preliminary data that show your project is feasible."[1]

PHS's comments further underscore the importance NIH attaches to a convincing showing of the feasibility of the proposed research through the presentation of preliminary results. Given this situation, it is hard to understand why PHS is reluctant to agree to a recommendation that either applicants be made aware of the importance reviewers place on such information or the discussion of preliminary results be made a formal, rather than optional, criterion.

8. Our recommendation is based on the specific finding that a high percentage of NIH reviewers reported only general knowledge of the subject of proposals they evaluated and a limited familiarity with the relevant literature. A number of reviewers responding to our survey commented on this issue—for example, "We vote, and our vote counts as much as the primary reviewer's. But on many proposals . . . I have no expertise at all." The recommendation is designed to address this problem by better using the expertise of study section members and ad hoc reviewers without radically altering the existing process. We have reworded the recommendation to clarify this point.

9. As we note in the report, both NSF and NEH retain the scores of reviewers, with appropriate safeguards for confidentiality. Nonetheless, it is possible that the alternative solution PHS proposes would be responsive

[1]Liane Reif-Lehrer, Writing a Successful Grant Application, 2nd ed. (Boston: Jones and Bartlett Publishers, 1989), p. 56.

to our recommendation. We are concerned, however, that a sample drawn to test for possible discrimination against minorities and women would not necessarily be adequate for other oversight and evaluation purposes, such as an audit of the accuracy of the computation of percentile scores. In addition, if PHS's proposal were a one-shot study, this would not provide data for continuing oversight and evaluation needs. A carefully selected, periodic (for example, annual) sample might be sufficient, however.

Comments From the National Science Foundation

Note: GAO comments supplementing those in the report text appear at the end of this appendix.

NATIONAL SCIENCE FOUNDATION
4201 WILSON BOULEVARD
ARLINGTON, VIRGINIA 22230

nsf

OFFICE OF THE DIRECTOR

March 24, 1994

Dr. Eleanor Chelimsky
Assistant Comptroller General
United States General Accounting Office
Washington, D.C., 20548

Dear Dr. Chelimsky:

Thank you for the opportunity to comment on your Office's draft report entitled Peer Review: Fairness in Federal Agency Grant Selection.

The Foundation welcomes the fact that the report finds no evidence for several negative stereotyped claims about peer involvement in Federal agency proposal review, such as supposed undue concentrations of reviewers from more prestigious institutions or certain regions of the nation. Through experience and self-study, we and other agencies have long been aware of the criticisms you examined, and have developed policies and practices to deal with them.

Indeed, your analysis indicates that the ". . . intrinsic quality of the proposals dominated the scoring process . . ." and that ". . . major factors influencing scores were those of the content of the proposals themselves, principally the questions raised and the designs." That should be the case, if the system is working properly.

We are deeply concerned, however, about certain statements and conclusions made in the report that are clearly at odds with available data. The small sample of actions relied upon by GAO for its study does not reflect a representative cross-section of all Foundation programs, or of our research programs as a whole; and the method chosen to calculate award rates led GAO to results that are directly contradicted by the actual rates achieved by the various types of applicants. To correct those portions of the report, I wish to provide the following information.

Award Rate Calculations

See comment 1.

We disagree strongly with the report's inaccurate depiction of NSF's award rates for women applicants, and with the implication that minority applicants generally receive lower scores and thus do not fare as well in receiving awards as non-minorities do. In

fact, our data clearly shows that NSF's award rates for women and minority applicants are roughly comparable to those for men.

The draft report incorrectly states that at NSF ". . . proposals from men are more than twice as likely to be funded as proposals from women." In fiscal year 1993, the Foundation as a whole made decisions on 29,860 competitively-reviewed proposals, and awarded 30% of them overall. A slightly higher percentage (32%) of the 5,922 proposals for which a woman was a Principal Investigator (PI) or co-PI were awarded.

Deleted.

Likewise, 27% of the 1,912 proposals having a minority PI or co-PI were funded in FY 1993, a little lower than the overall rate, but near the 30% and 31% rates of the prior four years. Examining proposals funded only through our research directorates (i.e., not including Education and Human Resources or Polar Programs accounts) shows a similar picture (See table in Attachment A).

GAO found lower-than-average reviewer ratings for the minority proposals it examined. But the report does not say how many of the total of 100 proposals that GAO examined were from minority applicants. Based on general patterns of NSF applications we presume it was five or fewer, across the five programs you chose from more than 200 that we operate.

As you correctly point out, NSF withholds from reviewers the information on race and ethnicity of applicants that we collect on a voluntary basis from proposers. It is not clear, therefore, what reasons -- aside from merit criteria such as proposal content and investigator experience -- might account for lower ratings. Indeed, your report finds ". . . no evidence to conclude one way or the other. . ." as to whether the ratings were biased, or fairly reflected merit factors.

See comment 2.

Perhaps the inexperience of new proposers accounts for lower ratings of some proposals from minority applicants. Proposals from minorities between FY 1988 and FY 1993 nearly doubled (see Attachment A). Any applicant's lack of experience in gauging the specific needs for frontier research in his or her field, and lack of experience in preparing proposals, would be very important factors in whether he or she receives an award. Our studies show that inexperienced applicants of any background who revise and resubmit their proposals overcome these impediments, and our policy is to help them do so.

Placing Reviewer Ratings in Context

Because of its focus on reviewer ratings, the report does not adequately describe the larger context in which agencies review and decide upon proposals.

First, NSF has long-standing policies, programs and outreach efforts that are expressly designed to encourage women, minorities and persons from other than elite universities to

apply and, other things being equal, to decide in their favor. The appendix to your report lists our various directed efforts, but the body of the report gives the Foundation very little recognition for having taken these initiatives over many years.

Second, the Foundation relies on the judgments of knowledgeable, well-trained program officers to integrate the written comments of reviewers -- which are frequently more informative and useful for decision-making than the ratings -- with general agency policies and, where they exist, with specific programmatic priorities.

At NSF, proposal decisions are based on the recommendations of program officers, not only on the reviewers' ratings and comments. We recruit experienced researchers and educators to be program officers, provide them with training, and call on them to:

- make thoughtful, well-justified proposal recommendations designed to move forward their field of research and/or accomplish stated educational objectives;
- take into account organizational policies and advice from the client community (through, for example, advisory committees and reports from professional societies);
- keep program content balanced, including where competing "camps" of proposers and reviewers may be involved;
- discount reviews that are uninformed, clearly biased, personally hostile, uninformative or otherwise not useful;
- communicate with the proposer to resolve unanswered issues raised by reviewers that could strongly influence the decision;
- communicate with potential applicants and declined applicants, to provide as much information and constructive advice as possible, particularly for proposers who are relatively "new to the system" ; and
- encourage talented people, particularly younger researchers, women and minorities to develop and submit proposals and to serve as reviewers.

We also expect program officers to monitor the progress of existing awards and to keep abreast of the intellectual frontiers of their programs by attending conferences, reading journals, and communicating with a wide range of persons in their field.

Reviewer comments and ratings are obviously quite important to the proposal review process. But they do not constitute the entire basis for making decisions. Our program officers must, and do, apply their own education, experience and judgment to the

merits of proposals and the contents of reviews, taking into account policy factors and program priorities. This creates the balance -- and the checks -- inherent in the process.

Third, NSF makes the peer review process as transparent as possible, by providing the reviews themselves, verbatim but unsigned, to the applicant, along with pertinent background information (such as the number of competing proposals and amount of funding available), and an explanation of the basis for the decision. This assists declined applicants who wish to revise their proposals and reapply.

Fourth, several internal control and oversight mechanisms exist to keep the review system working openly and fairly. Let me describe seven safeguards that the Foundation has built into the system:

See comment 3.

Now p. 62.

1. A body of conflicts-of-interests rules has been developed that are designed to surface and resolve conflicts situations. (Note that contrary to the statements on pp. 3-20 and 3-21 of the draft report, NSF's regulations at 45 CFR 681.21 do cover all three examples mentioned: personal friendship, past collaboration, and teacher-pupil relationships).

2. Program officer recommendations (whether to award or decline a proposal) are reviewed at least by the person at the next higher organizational level, who must concur in, or reject, each one;

3. Recommendations involving sizable amounts of funds, new program efforts, etc., are also reviewed further up the organization and, in some cases, by the National Science Board (the Foundation's governing body), before the award can be made;

4. Declined applicants are informed as to how to have their original proposal reconsidered at two successively higher levels if they believe it was unfairly handled;

5. Visiting committees of knowledgeable persons from the various clientele communities closely examine each program every three years for, among other things, fairness and balance in decisions;

6. Several hundred proposal files chosen at random are examined by NSF's Office of Inspector General each year, for conformance with procedure and policy; and,

7. The Foundation, for its own management purposes, has conducted surveys and other studies of the effectiveness and outcomes of the review system.

As your report points out, ". . . the findings from such (agency) studies are hardly self-serving. . .". In the case of NSF, they have not only illuminated the system but led to many changes over the years.

Finally, the draft report does not adequately convey NSF's long history of responsiveness in policy and practice to concerns about the review system -- such as the set of changes made in 1990, based in large part on self-study, to establish Small Grants for Exploratory Research, to clarify our process for reconsideration of proposal decisions, and to permit proposers to suggest reviewers and persons not to review their proposals.

Deleted.

Additional technical comments on the report are contained in Attachment B.

Comments on Report Recommendations

See comment 4.

We agree that we should monitor more closely our efforts to include more women, minorities and younger researchers in the review process, especially in programs that do not use panels. We view such efforts as important for providing diverse viewpoints on the merits of the scientific or educational content of proposals, not as mere representation for its own sake. We should note, however, that legal considerations regarding privacy do not allow us to ask about gender or ethnicity of reviewers, or to keep that information in our databases.

We also note, and will look for ways to address, the need to identify any unwritten or informal decision rules commonly used by reviewers, and to better calibrate reviewer ratings through examples and other information. Proposers and reviewers should all have an accurate and similar understanding of the review criteria.

See comment 5.

We do not agree, however, that changing the review form to provide separate scores for each criterion would improve "fairness"; indeed, moving in that direction could easily lead to reliance on merely averaging reviewer scores -- to the point of making meaningless arithmetical distinctions, and relying less on reviewer comments, which are more important than scores alone.

See comment 6.

Nor are "blinded" proposal reviews worth further testing. Efforts to conduct "blinded" reviews often falter because reviewers attempt to guess the identity of the applicant, sometimes mistakenly. More importantly, the capability of the proposer to do the work -- based on qualifications, potential and experience -- is an essential factor in the award decision. Reviewers must have that information to do their work conscientiously.

During my first months as NSF director, I have been impressed with the overall quality of the proposal review system, with the awareness of the staff about the community's concerns, and with the dedication to operating a fair system that results in effective investments of taxpayer funds -- despite handling proposal workloads half again higher than a decade ago, with virtually no increase in staffing levels.

The Foundation has a history of making changes in the proposal decisionmaking process when systematic problems arise. I can assure you that during my tenure we will continue to pay close attention to the workings of the review system and change it again if necessary.

Sincerely,

Neal Lane

Neal Lane
Director

The following are GAO's comments on the March 24, 1994, NSF letter.

GAO Comments

1. We note in numerous places that the data we report represent a sample that was selected to have approximately equal numbers of funded and nonfunded applications. As we note in chapter 1, we employed this strategy so that we could compare successful and unsuccessful applications on a range of dimensions. We set out not to answer the question of what proportions of women and minorities were funded but, rather, what factors were related to awards and declinations. Because of this sampling strategy, we neither claim nor imply that the success rates reported in chapter 4 represent the overall success rates for women across all NSF programs—the study was not designed to generate that estimate.

In addition, it is not clear that the data NSF cites are comparable to those we analyzed. We considered only whether the primary investigator was male or female. In contrast, the analysis cited in the NSF letter also includes cases in which women are co-investigators. But including co-investigators in the calculations could have the effect of inflating the percentage of successful women applicants.

The issue related to success by minority applicants is more subtle. We did not report a direct relationship between minority status and funding success. However, we did note that the primary factor affecting whether a proposal was funded was the peer review score it achieved, and we noted that applications from minorities generally got lower scores. We suggested only that this could result in minorities faring less well than whites, and indeed the data NSF cites, even though they include minority co-investigators, do show a slightly lower success rate for minorities.

2. We agree with NSF that the somewhat lower scores for applications submitted by minority investigators may reflect a number of factors, including inexperience, and indicated this in the draft NSF reviewed. It should be noted that in our analyses we did control for the professional age of the applicant—which logically should be related to professional experience—and minorities still got lower scores.

3. The section of the regulation NSF cites applies to program officers, not peer reviewers. Section 681.25, which does apply to reviewers, is not explicit about what conflicts are covered, although it does permit program officers to use the list in section 681.21 "as a guide in responding to reviewer questions."

4. We are not aware of any legal considerations that would prevent NSF from collecting and retaining information on the gender or ethnicity of reviewers, although reviewers cannot be compelled to provide such data. In fact, the data sets we received from NSF specifically included data on the gender of each reviewer.

5. We did not recommend numerical scoring; to clarify this point, we have changed the wording of the recommendation. Furthermore, we do not believe our recommendation would lead to the problem of "averaging" alluded to by NSF. In fact, we specifically include a separate "overall" rating as part of the recommendation. We are suggesting a review format that would ensure that all elements of each proposal are at least considered by reviewers before an overall evaluation is reached. Nothing would require reviewers to give any particular weight to these elements in reaching an overall rating, so the issue of averaging should not come up.

6. NSF misinterprets our recommendation. We do not advocate using blind reviews as the method for evaluating proposals. Rather, we suggest using them in a sample of cases to compare with nonblind reviews as a check on possible discrimination.

Comments From the National Endowment for the Humanities

Note: GAO comments supplementing those in the report text appear at the end of this appendix.

NATIONAL ENDOWMENT FOR THE HUMANITIES

WASHINGTON, D.C. 20506

March 15, 1994

MEMORANDUM

To: Dan Rodriguez
Project Manager
Program Evaluation and
Methodology Division
General Accounting Office

From: Donald Gibson
Acting Deputy Chairman
National Endowment for the Humanities

Subject: NEH's Comments on GAO's Draft Report on the Peer Review Systems of NEH, NSF, and NIH

Thank you for this opportunity to comment on the draft of GAO's report "PEER REVIEW: Fairness in Federal Agency Grant Selection Processes." The report, as directed by the Senate Committee on Governmental Relations, presents the results of GAO's study of potential and actual weaknesses with peer review in federal funding. We believe that this study--which examines the peer review systems of the National Endowment for the Humanities (NEH), the National Science Foundation (NSF), and the National Institutes of Health (NIH)--focuses on many important issues that relate to the equity of peer review as a method for awarding federal funds. NEH shares these concerns and works hard to maintain its reputation for operating a fair and equitable system for reviewing grant proposals in the humanities. We think that GAO project staff have done an estimable job of describing, surveying, and analyzing the Endowment's process of peer review.

We found the report's recommendations to be thoughtful and useful suggestions for helping NEH to preserve the high quality of its peer review system. We will be exploring the possibility of implementing a number of these recommendations including recruiting more junior faculty to serve as NEH reviewers.

Attached are NEH's responses to some of the points raised in the draft report that relate to our peer review system. This material is organized as follows: our general comments and overall impressions of the study; our comments on and responses to the report's major findings and recommendations; and an item-by-item discussion of other references to the NEH review process

in the text of the report that we think need to be clarified or explained more fully. These responses contain a number of suggestions for revising selected report language that, if adopted, we think would strengthen the final draft of the report.

If you have any questions about NEH's comments on GAO's report--"PEER REVIEW: Fairness in Federal Agency Grant Selection Processes"--please do not hesitate to contact me.

Attachment

NATIONAL ENDOWMENT FOR THE HUMANITIES

I. General Comments on the Draft Report

With the exceptions noted below, the National Endowment for the Humanities is pleased with the draft report's overall characterization of our review system. We think that the report validates the widely held view in the humanities community that our peer review system is fair, judicious, and objective. In our reading of the report, peer review at NEH is depicted as unbiased in the way it functions to identify grant proposals in the humanities as candidates for federal funding. We also note that a vast majority of the reviewers who participated in the study concluded that while peer review may not be a perfect system and would benefit from some minor reforms, it was a system that "was fair or at least the best available." Moreover, although the report contains many references to "potential" problems that may arise during peer review, "weaknesses" in some procedures, and "possibilities" for abuses to occur, relatively little actual evidence of problems concerning the operation of NEH's review system was found. In this regard, we thought that the charts that make up Appendix I: "Framework Identifying Potential Weaknesses and Agency Actions to Address Them" present a thorough and useful compilation of the policies and procedures NEH currently has in place to guard against possible inequities in its review system. We would add that the Endowment's staff and leadership are dedicated and fully committed to maintaining the integrity of our system.

See comment 1.

Although we think, as we have said, that the report was generally well done and that it provides an overall favorable assessment of NEH's review system, we also have a number of problems with the study. First, we must say that we are somewhat at a loss to understand why NEH was included in this study, which is focused primarily on the conduct of peer review in the sciences. Indeed, the report's basic premise, as stated on pp. 1-6/7, is that "(m)any critics believe that peer review is a system in which a select group of scientists [emphasis added] from a small number of elite universities repeatedly decide to fund each other's research while waving the flag of merit review to justify these decisions." In addition, the report asserts, on p. ES-1, that "there has been a long history of controversy about how peer review is actually practiced, particularly regarding the fairness of peer review processes." While we cannot speak to the record of criticism of the peer review systems of NSF and NIH, these assertions are not descriptive of NEH's review system. We are not aware and the report does not present any evidence of criticism of our review procedures; an examination of the report's bibliography, for instance, indicates that the published sources used in the study are concerned almost entirely, if not

Now p. 13.

Now p. 1.

exclusively, with peer review in the sciences.

We are compelled to point out that NEH's review system is highly regarded in the humanities community and that it has functioned remarkably well over the almost thirty-year history of the agency. This is not to say, of course, that we do not occasionally receive complaints--we do. But these criticisms normally come from applicants whose proposals were rejected and who will question specific verdicts rendered by the system, not the system itself. Indeed, we have received many testimonials to our review system over the years--from both inside and outside the humanities community. Peer review at NEH is healthy and is not in need of major reform. We are confident that the overwhelming majority of our applicants and of those who have participated in the system would agree with this assessment.

See comment 2.

Another general problem with the design of the study is that it focuses exclusively on peer review. We understand that this is the inevitable result of the mandate that GAO received from the Senate Committee on Governmental Relations, but it also has the consequence of obscuring the true nature of NEH's review system. NEH has a multi-tiered review process: Applications are scrutinized not only by specialist reviewers and panels, but also by NEH staff, the National Council on the Humanities, and the Chairman of the Endowment, who is charged by legislative statute with making the final decision on whether or not to fund. This process ensures that each application is exposed to a spectrum of perspectives. Thus, while peer review is the most important step in the system, it is neither the only nor the final determinent of an application's status.

See comment 3.

The report also generalizes about the review procedures of the agency as a whole but only a small number of our more than twenty-five programs were sampled. The study concentrates on applications from individuals and fails to consider the high percentage of institutional applications we receive; annually, over one-half of our grants and about 94 percent of our funding are for institutional projects in the humanities. Many of our programs have different missions, purposes, and procedures that are designed to meet their varied objectives in the humanities. Thus, some of the study's generalizations are not applicable to nor reflective of all of NEH's grant programs.

In addition to the basic premises and the design of the study, we also question the accuracy and implications of some of the study's specific findings and conclusions. We are troubled that other readers of the report may draw false impressions about peer review at NEH from this material. In the following two sections of this paper, we discuss in detail some of the problems we have identified in the study. We offer these comments as examples of constructive ways the report can be improved that will increase its usefulness as a reliable source of information

about the peer review component of the evaluation system of the National Endowment for the Humanities.

II. NEH's Responses to the Report's Principal Findings and Recommendations

The report's major findings and recommendations are itemized in both the opening "Executive Summary" section and in the concluding chapter of the text. We have comments to make about each of the findings or conclusions that are applicable to the Endowment. We also will comment on each of the report's four general recommendations and the two additional recommendations that are directed specifically to NEH.

A. Comments on the Report's Principal Findings

1. Selection of peer reviewers

Now p. 3.

Report, pp. ES-4/5: "At all three agencies, large majorities of reviewers reported that their own scholarly expertise was at least in the general area of the proposals they reviewed. However, reviewers often did not work on questions related to the application, nor could they cite much of the literature. . . ."

See comment 4.

NEH Comment: This finding needs to be clarified in regard to NEH. We think that all of the reviewers that we select to serve in our review system have the requisite level of knowledge and expertise that is needed to evaluate fairly and competently the applications we have assigned to them. In the humanities, unlike, perhaps, in the sciences, it is not necessary for every reviewer, particularly every panelist, to have worked directly "on questions related to the application," as this passage of the report implies. In fact, we strive to have generalists as reviewers on some of our panels to help us gauge the broad significance to the humanities of specialized proposals. Indeed, the report acknowledges the inherent breadth of our review system by observing on p. 2-12 that NEH panels "represent a very broad spectrum of the humanities because it is a small agency that supports numerous disciplines, and the disciplines themselves cover world-wide topics and all periods of civilization."

Now p. 32.

Now p. 3.

Report, p. ES-5: "While reviewers were broadly similar to applicants on a number of factors (such as the region of the U.S. where they were employed), young scholars and those with lower academic rank were consistently under-represented among reviewers at all three agencies. At all three

agencies, panels included at least the same proportion of women as among applicants. . . ."

See comment 5.

NEH Comment: We disagree with the report's basic assumption, as expressed in this and many other passages, that reviewers should closely match the demographic profile of the applicant pool. We think that it is in the best interest of our applicants and for the advancement of the humanities for NEH to select only the most competent educators, scholars, and professionals in the humanities to serve as our peer reviewers, regardless of those reviewers' demographic characteristics. Within this context, we strive for and are committed to having as diverse a group of reviewers as possible. To help us in this task, we employ a computerized system for cataloguing and retrieving the names, addresses, and qualifications of more than 21,000 prospective panelists and reviewers. The system greatly facilitates the job of selecting qualified individuals to serve as NEH panelists and reviewers while providing appropriately for geographic and cultural diversity.

See comment 6.

While this passage acknowledges and commends the geographic and gender diversity of our panels, it also suggests that we under-utilize the services of younger or junior scholars. We would point out that many junior faculty members annually serve the Endowment as panelists, particularly in programs like the Fellowships for College Teachers and Independent Scholars and the Study Grants for College Teachers programs in our Fellowships and Seminars division. The Endowment also regularly recruits faculty at community colleges to serve on panels in programs that receive a number of applications from those institutions such as those in the Education Division. Morevover, it has been our experience that junior scholars are often some of our more conscientious and insightful reviewers.

2. Factors related to scoring of proposals

Now p. 3.

Report, p. ES-5: "GAO observed unwritten decision rules being applied by reviewers at all three agencies. This can give applicants who have served as reviewers an unfair advantage over those who have not."

See comment 7.

NEH Comment: We think it is necessary to point out for the record that we do provide all of our panelists--both sitting panels and mail panels--with written instructions on how they are to evaluate applications. All of the governing criteria for evaluating proposals are enumerated in these materials. The beginning of each panel meeting, we also reiterate for panelists the evaluation criteria they are to use during their proceedings. We think that there is a limit, however, to how much formal guidance should be

provided to our sitting panelists. It is important to note that NEH panel meetings are deliberative by design: They are based on the idea that through free discussion and debate, panelists will arrive at their decisions regarding the relative merits of the applications that are under review. Such free-ranging discussions may appear to observers who are not familiar with the workings of the NEH review process as "unwritten decision rules," but we can assure you that these panel deliberations are within the parameters of the instructions that we have provided to the panelists. We would not want to do anything that would stifle the collegial and deliberative nature of this process as it exists now.

See comment 8.

We categorically dispute the report's contention that applicants who have served as reviewers have "an unfair advantage over those who have not." If an applicant has first-hand knowledge of our review procedures, then that applicant will also know that the system is fair and rigorous and that only the highest quality and most significant grant proposals in the humanities are likely to negotiate the system successfully and be approved for funding. Thus, mere knowledge of our system is not going to benefit an applicant vis-a-vis other applicants if that applicant does not have a a competitive application.

Now p. 4.

Report, p. ES-6: "At all three agencies reviewers tended to give better scores to proposals from applicants perceived to have stronger academic reputations, leaving unknown applicants at a disadvantage. In addition, at both NSF and NEH, reviewers gave higher scores to applicants they knew than to those they did not know."

See comment 9.

NEH Comment: In response to the first part of this passage, we would point out that the NEH review system is designed to focus on the quality and significance of the humanities proposal that is submitted to us for consideration, regardless of who the applicant is. We fund applications in the humanities, we do not fund academic reputations: Applicants that are well-known in the humanities must, like all other applicants, develop a significant project and submit a quality application if they expect to receive serious consideration for funding. We would also submit that it is probable that, in the aggregate, applicants with established track records (that is, those with "stronger academic reputations") will naturally tend to develop the most significant proposals and hence to receive the highest scores from our review system. But we see this as a testament to our system's ability to identify the most significant projects rather than as a reflection of the applicant's "reputation" as implied by the report.

See comment 10.

Similarly, it is also probably true that the top practitioners of the humanities are better "known" by reviewers because of their prominence. This is a natural outgrowth of their intellectual stature and not the fact that they are "known" in a personal sense by reviewers.

Now p. 4.

Report, p. ES-6: "These results could suggest that experienced, well-known, white, male scholars write better proposals than others, or know the rules and norms for proposal-writing better, but they also could be viewed as supporting some of the equity concerns presented by critics and raise at least the possibility of bias in the scoring process. However, GAO's statistical analysis left much of the variation in scores unexplained, suggesting that the intrinsic quality of the proposals dominated the scoring process."

See comment 11.

NEH Comment: This finding of the report, like a number of other conclusions, is based on the hypothetical possibility of bias rather than on any objective evidence of inequity. This is a reflection, perhaps, of the design of the study, which is to try to identify potential weaknesses in peer review systems where bias may occur. As this passage also observes, however, funding decisions at NEH may actually reflect "the intrinsic quality of the proposals" rather than some perceived bias for or against certain kinds of applicants.

3. Factors related to decisions on funding

Now p. 4.

Report, p. ES-7: ". . . at NEH, funding decisions were strongly related to the dollar amount requested. . . . Generally, the likelihood that a project would be funded depended on score, but beyond some point, the odds of funding decreased sharply for proposals with worse scores and higher requested amounts."

See comment 12.

NEH Comment: We are somewhat puzzled by this finding. On the one hand, as stewards of public funds we think that the federal government has a responsibility to examine budgets closely. On the other hand, there is no direct correlation that we know of between the amount of funding requested by an applicant and the likelihood that the applicant will receive funding. The study's survey results may appear to show a connection between reviewer scores and amount requested, but we submit that there is no evidence to suggest that there is a causal relationship between these two variables. Indeed, throughout the Endowment, some of the more complex and costly projects that apply for grant support receive some of the highest grades from reviewers.

Now p. 4.

See comment 13.

Now p. 4.

See comment 14.

B. Comments on the Report's General Recommendations

Recommendation 1, p. ES-8: "address the exclusion of young, junior faculty as reviewers by targeted outreach efforts, at least on an experimental basis"

NEH Comment: This is a useful suggestion that will help to underscore our continuing outreach efforts in all phases of our operations. We are committed to making our review system as open and as inclusive as possible. We suggest, however, that the word "exclusion" be changed in the wording of the recommendation: NEH does not "exclude" junior faculty from serving in our review system. Indeed, as we noted previously, such faculty members regularly serve on panels in programs throughout the Endowment. Thus, we think that "under-utilization" would be a more appropriate and more accurate term to use than "exclusion" in the wording of this recommendation.

Recommendation 2, p. ES-8: "increase monitoring of discrimination, including conducting tests using 'blind' reviews to assure that sex, race, and ethnic discrimination are not affecting scores given by peer reviewers" (p. ES-8)

NEH Comment: In response to the report's recommendation, we are considering the feasibility of experimenting with "blind" reviews in at least one of our programs. "Blind" reviews would have only limited applicability to NEH: Because the review process in many, if not most, of our grant programs is predicated on reviewers having at least some knowledge of the applicant's academic, scholarly, and professional background and accomplishments in the humanities because this indicates ability to carry out the project successfully. An NEH-wide "blind" review system would not be possible and would not be a responsible method of awarding public funds.

We would like to point out that we are ever-vigilant to make sure that "discrimination" on the basis of any factor other than the quality of the proposal under review does not enter into our application evaluation process. Indeed, our review system has an exemplary track record and a hard-earned reputation in the humanities community for being fair and unbiased. The multi-tiered structure of our review system ensures that each application is exposed to a spectrum of perspectives and helps to guard against the possibility that a panelist or group of panelists may be biased for or against an applicant on the basis of gender, race, ethnicity, or any other factor.

It should also be noted that the NEH review system is currently "blind" to a degree--our standard application cover sheet does not contain information on the applicant's race, age, or ethnic affiliation. In some instances, it is also unlikely that reviewers would know the gender of the applicant.

Now p. 4.

Recommendation 3, p. ES-8: "address the lack [of] clarity in the application of scoring criteria by considering use of a scoring system in which proposals are rated separately on a number of criteria as well as in a summary score"

See comment 15.

NEH Comment: This recommendation would have limited applicability in NEH's grant programs. As we have already noted, the Endowment's panel process is primarily deliberative in nature and does not lend itself to a strictly numerical scoring system that may be more appropriate, perhaps, for the review of proposals in the sciences. At our panel meetings, panelists are encouraged to discuss and to evaluate an application in its entirety, and not just grade the application's component parts.

We would mention that one Endowment program has already experimented with using a weighted scoring system: In FY 1993, the first stage of review of applications to our new Dissertation Grants program involved the use of 25 panels that were polled by mail for their evaluations of the approximately 1,500 applications received in the competition according to a specific set of criteria. While we found this to be a useful and efficacious method for obtaining preliminary assessments on a high volume of applications, we do not think such an approach would be a responsible review procedure to adopt in our other, more established programs.

We also think the humanities community would oppose any effort to implement a strictly weighted, numerical system of evaluating grant applications in all Endowment programs.

Now p. 4.

Recommendation 4, p. ES-8: "identify any commonly used unwritten or informal decision rules that are applied by peer reviewers, and where feasible formalize them, or at least inform applicants of their importance"

See comment 16.

NEH Comment: We believe that on the whole this a good suggestion: We will try to do a more effective job both of standardizing our instructions to reviewers and of informing applicants how reviewers arrive at their evaluations. As we have stated earlier in this paper, however, there is a limit to the extent to which we may be able to "formalize" the steps that panelists take to arrive at their recommendations. We would be loath to implement any revision in our procedures that would compromise or obstruct

the deliberative nature of our panel process.

C. Comments on the Report's Specific Recommendations for NEH

Now p. 5.

Recommendation 1, p. ES-9: "improve evaluation and oversight by retaining data on the race of applicants"

See comment 17.

NEH Comment: We are reluctant to begin collecting data on the race of our applicants. In the absence of any evidence that race plays a role in our review process, which the report does not provide, we do not see the need to collect such data. We also think that many individuals and institutions in the humanities would object to supplying these data.

If we were to collect information on the race of applicants, such data would have limited usefulness. Many of the grant proposals we receive--especially in the Research, Education, and Preservation and Access divisions--are essentially institutional applications that involve more than one scholar or project director; hence, it would be difficult if not impossible to assign a definitive "race" classification to these applications.

Now p. 5.

Recommendation 2, p. ES-9: "increase efforts to calibrate proposal ratings among reviewers through information provided in advance and discussions of examples at the convening of panels"

See comment 18.

NEH Comment: As we have already stated, we will try to standardize the instructions and information we send to panelists as much as possible. But, as we have said, we also do not want to be too proscriptive and to restrain the deliberative process that lies at the heart of our panel system.

We also point out that this passage of the report contains misleading information: In most of our programs, we do in fact discuss selected applications at the opening of panel meetings so that panelists can develop a sense of the relative merits of the applications they will be evaluating.

The following are GAO's comments on the March 15, 1994, NEH letter.

GAO Comments

1. NEH was included in the study precisely because the congressional requester did not want to limit the scope of our work only to the sciences. We have tried to change language in the report that might have suggested inadvertently that it applies only to the sciences. Moreover, we have added references to articles specifically criticizing peer review at NEH.

2. In the report, we explicitly recognize the role of staff and councils in making funding decisions at all three agencies. Indeed, the whole analysis in chapter 4 rests on the recognition that peer reviewers are not the ultimate decisionmakers at any of the agencies we studied.

3. We make clear in chapter 1 which NEH programs we examined, and we make no effort to generalize our findings to other NEH programs. However, we have added language to clarify that we examined only individual, rather than institutional, applications in our review.

4. We do not say or imply that every reviewer should have worked on the same question raised in the proposal and, as NEH acknowledges, specifically note the underlying philosophy of panel construction at the agency.

5. We did not assume that reviewers necessarily should mirror applicants in terms of demographic characteristics. Our effort was, rather, to test against actual data complaints made about underrepresentation of various groups in the review process. By comparing reviewers with the applicant pool, we were able to show that differences in the numbers of reviewers from various demographic groups, regions, and types of institutions largely flow from differences in the numbers of such persons among active scholars. All such comparisons can be misinterpreted; had we not made this comparison, readers might have understood us as implying that each group should be equally represented among reviewers.

6. NEH takes issue with our finding that younger scholars are underrepresented among its reviewers. However, the agency later says it finds our recommendation to address this issue "useful." We believe our data accurately represent the situation.

7. We agree that it would be unfortunate to stifle discussion of proposals at panel meetings. Our point is simply to note what we observed at panel

meetings at all three agencies: some issues tended to emerge quite consistently as part of panel discussions, and these issues were not necessarily communicated to applicants beforehand. We have termed these issues "unwritten decision rules." In chapter 3, we list a number of specific factors that came up during our observations of NEH panels.

8. We do not claim that prior service as a reviewer, in and of itself, confers an advantage on the applicant. What we do argue in chapter 3 is that if there are unwritten rules that affect how proposals are scored, then applicants with experience as reviewers would be aware of them and would be more likely to address those issues in their proposals. All other things being equal, this could give them a "leg up" in securing a better score and subsequently being awarded funding.

9. We have deleted the language suggesting that unknown applicants are at a disadvantage. Our point was to describe the statistical relationships we found, and therefore we have retained the finding that those with stronger perceived track records score better than others.

10. Actually, we did ask reviewers whether they "knew" applicants in a personal sense and we found that, in the aggregate, NEH reviewers did give better scores to applicants whom they knew personally. We agree that top practitioners of the humanities may be more widely known than others because of their professional prominence, as we discuss in chapter 3.

11. As we note in the passage NEH cites, our statistical findings on the relationships between scores and a number of other factors are open to several interpretations. We have tried to present the range of plausible interpretations wherever possible. We could not, of course, examine the motives of individual reviewers, nor did we try.

12. NEH claims "there is no direct correlation" between the amount requested by an applicant and the likelihood of funding. However, correlation is precisely what we found in the data. We do not claim that amount requested is consciously employed as a screening variable in making funding decisions; it may be that the influence is more subtle, or that—as with all statistical analyses—our findings are the result of chance alone (though the probability of chance is small). We are simply reporting what the data say.

13. We have changed "exclusion" to "underrepresentation."

14. NEH misunderstands the point of our recommendation. We do not advocate using blind reviews as the method for awarding public funds. Rather, we suggest using them in a sample of cases to compare with nonblind reviews as a check on possible discrimination. We recognize that NEH attempts to remove race and gender as possible sources of bias by excluding this information from the application cover sheets distributed to reviewers. However, as we note in chapter 3, it is often possible for reviewers to learn, or in any case guess, the race and gender of an applicant from other available information. In fact, at one NEH panel we attended, a reviewer specifically announced to other panel members the deduction that one applicant was a minority female.

15. We did not recommend numerical scoring; to clarify this point, we have changed the wording of the recommendation. Furthermore, we do not believe our recommendation would discourage the discussion of proposals. In fact, we observed at panel meetings that discussions of individual proposals sometimes focused on one or two aspects, while neglecting others. What we are suggesting is a review format that would ensure that reviewers at least consider all elements of each proposal before an overall evaluation is reached. Curiously, NEH concedes that its own experience with a weighted system of ratings was successful but argues that such a system works only for the preliminary consideration of proposals. NEH does not explain how it reached this conclusion.

16. We agree that there are practical limits on NEH's ability to formalize the review process but doubt that those limits have yet been reached.

17. The absence of evidence in our report that race plays a role in NEH's review process reflects precisely the lack of race data in the agency's files that our recommendation is designed to remedy. It means not that we found no effect of race on decisions but, rather, that we could not study the issue because the agency lacked the necessary data. Given the relationships we did find at NSF, where data were available, we believe NEH ought to collect the information it would need to monitor this area. Furthermore, our recommendation is addressed only to applications for funding of the work of individuals, not institutions; we did not study institutional grants.

18. We are happy that NEH will try to improve the standardization of instructions. However, we must also note that in our observation of NEH panel meetings, we did not see the discussion of selected applications,

with a view toward calibration of rating standards, alluded to in NEH's comments.

Reviewers of This Report

John Aherne
Sigma Xi, The Scientific Research Society
Research Triangle Park, N.C.

Stephen Cole
State University of New York
Stonybrook, N.Y.

Marcel La Follette
Center for International Science, Technology, and Policy
George Washington University
Washington, D.C.

James Savage
University of Virginia
Charlottesville, Va.

Michael Scriven
Evaluation and Development Group
Inverness, Calif.

Louise Tilly
Committee on Historical Studies
New School for Social Research
New York, N.Y.

Carol Weiss
Graduate School of Higher Education
Harvard University
Cambridge, Mass.

Rosalyn Yalow
Veterans Affairs Medical Center
Bronx, N.Y.

Major Contributors to This Report

Program Evaluation and Methodology Division

Patrick G. Grasso, Assistant Director
Daniel G. Rodriguez, Project Manager
Elaine L. Vaurio, Social Science Analyst
James Joslin, Social Science Analyst
Harry M. Conley III, Sampling Consultant
Venkareddy Chennareddy, Referencer
Penny Pickett, Supervisory Reports Analyst
Robert Fiorentine, Project Adviser

Bibliography

Carter, Grace M. "A Review of the Literature Concerning the NSF Peer Review System." Prepared for the Division of Policy Research and Analysis, National Science Foundation, 1986.

Chubin, Daryl E., and Edward J. Hackett. Peerless Science: Peer Review in U.S. Science Policy. Albany, N.Y.: SUNY Press, 1990.

Clark, Allan. "Luck, Merit, and Peer Review." Science, 215 (1982) 346-48.

Cole, Jonathan R., and Stephen Cole, Peer Review in the NSF: Phase 2. Washington, D.C.: National Academy of Sciences, 1981.

Cole, Stephen. Making Science. Cambridge, Mass.: Harvard University Press, 1992.

Cole, Stephen, and Jonathan R. Cole. "Chance and Consensus in Peer Review." Science, 214 (1981), 881-86.

Cole, Stephen, and Jonathan R. Cole. Peer Review in the NSF: Phase Two. Washington, D.C.: National Academy of Sciences, 1981.

Gillespie, Gilbert W., Jr., Daryl E. Chubin, and George M. Kurzon. "Experience with the NIH Peer Review: Researchers' Cynicism and Desire for Change." Science, Technology, and Human Values, 10 (1985), 44-54.

Gourman, Jack. The Gourman Report. Los Angeles, Calif.: National Education Standards, 1989.

Greenberg, Daniel S. The Politics of Pure Science. New York: New American Library, 1968.

Jones, Lyle, V. An Assessment of Research-Doctorate Programs in the United States, vols. 1-5. Washington, D.C.: National Academy of Sciences, 1982.

Martino, Joseph P. Science Funding. New Brunswick, N.J.: Transaction Publishers, 1992.

Merton, Robert K. "The Matthew Effect in Science." The Sociology of Science. Chicago, Ill.: University of Chicago Press, 1973.

National Institutes of Health, NIH Grants Peer Review Study Team. Grants Peer Review: Report to the Director, NIH Phase I. Washington, D.C.: 1976.

National Science Foundation. Grants for Research and Education in Science and Engineering (GRESE), report 90-77. Washington, D.C.: 1990.

National Science Foundation Advisory Committee on Merit Review. Final Report. Washington, D.C.: 1986.

National Science Foundation, Program Evaluation Staff. Proposal Review at NSF: Perceptions of Principal Investigators, report 88-4. Washington, D.C.: February 1988.

Roy, Rustum. "Funding Science: The Real Defects of Peer Review and the Alternative to It." Science, Technology, and Human Values, 10 (1985), 73-78.

Savage, James D. "The Distribution of Academic Earmarks in the Federal Government's Appropriation Bills, FY 1980-89." Working Paper 89-5. University of California at Berkeley, Institute of Governmental Studies, Berkeley, Calif., 1989.

Shapley, Deborah and Rustum Roy. Lost at the Frontier: U.S. Science and Technology Policy Adrift. Philadelphia, Pa.: ISI Press, 1985.

Sigma Xi, The Scientific Research Society. "A New Agenda for Science," preliminary report, New Haven, Conn., 1986.

U.S. General Accounting Office. Better Accountability Procedures Needed in NSF and NIH Research Grant Systems, GAO/PAD-81-29. Washington, D.C.: 1981.

U.S. General Accounting Office. University Funding: Information on the Role of Peer Review at NSF and NIH, GAO/RCED-87-97FS. Washington, D.C.: 1987.

U.S. General Accounting Office. University Funding: Patterns of Distribution of Federal Research Funds to Universities, GAO/RCED-87-67BR. Washington, D.C.: 1987.

Yalow, Rosalyn S. "Is Subterfuge Consistent With Good Science?" Bulletin of Science Technology and Society, 2 (1982), 401-4.

Zuckerman, Harriet, Jonathon R. Cole, and John T. Brues. The Outer Circle: Women in the Scientific Community. New Haven: Yale University Press, 1992.

Ordering Information

The first copy of each GAO report and testimony is free. Additional copies are $2 each. Orders should be sent to the following address, accompanied by a check or money order made out to the Superintendent of Documents, when necessary. Orders for 100 or more copies to be mailed to a single address are discounted 25 percent.

Orders by mail:

U.S. General Accounting Office
P.O. Box 6015
Gaithersburg, MD 20884-6015

or visit:

Room 1100
700 4th St. NW (corner of 4th and G Sts. NW)
U.S. General Accounting Office
Washington, DC

Orders may also be placed by calling (202) 512-6000 or by using fax number (301) 258-4066.

Each day, GAO issues a list of newly available reports and testimony. To receive facsimile copies of the daily list, or any listing from the past 30 days, please call (301) 258-4097 using a touchtone phone. A recorded menu will provide information on how to obtain these listings.

United States
General Accounting Office
Washington, D.C. 20548-0001

Bulk Mail
Postage & Fees Paid
GAO
Permit No. G100

Official Business
Penalty for Private Use $300

Address Correction Requested