National Science Foundation

PHASE ONE OF A STUDY

The National Research Council

INSTITUTE OF MEDICINE · NATIONAL ACADEMY OF ENGINEERING · NATIONAL ACADEMY OF SCIENCES

PREPARED FOR THE COMMITTEE ON SCIENCE AND PUBLIC POLICY OF THE NATIONAL ACADEMY OF SCIENCES

Peer Review in the National Science Foundation

PHASE ONE OF A STUDY

STEPHEN COLE
Professor of Sociology,
State University of New York at Stony Brook

LEONARD RUBIN
Consultant,
National Academy of Sciences

JONATHAN R. COLE
Professor of Sociology,
Columbia University

NATIONAL ACADEMY OF SCIENCES
Washington, D.C. 1978

Library of Congress Catalog Card Number: 78-55184

International Standard Book Number 0-309-02788-8

Available from

Office of Publications
National Academy of Sciences
2101 Constitution Avenue, N.W.
Washington, D.C. 20418

Printed in the United States of America

Foreword

This report by the Committee on Science and Public Policy of the National Academy of Sciences addresses a matter of central concern to scientists and to the general public: How are the judgments made that determine which specific basic research projects and investigators shall be supported with the funds allocated to such purposes by the Congress? The National Science Foundation is charged with assuring the continuing strength of national scientific endeavor. Accordingly, it is the responsibility of the Foundation to determine which areas of science should be supported and in what relative amounts. Within each area, the Foundation must identify those research projects that offer greatest opportunity either for advancing understanding or for subsequent application. The principal mechanism utilized by the Foundation to this end is the "peer review process."

It is the purpose of this report to describe and examine the operation of that process in light of the above purposes. I am grateful indeed to the Committee on Science and Public Policy for this effort, which should serve science and the nation well.

PHILIP HANDLER
President

Preface of the Committee on Science and Public Policy

Both the scientific community and the public at large want to be sure that innovative creative research is supported as effectively as available funds permit. They believe that support should be provided not only for the active leaders of the major scientific fields but for talented young researchers early in their careers. Most would also argue that scientists whose productivity and originality are declining should receive less support. In any event, the public has a right to know whether its monies are wisely spent, whether the funds available are in fact distributed to support research of the highest quality. It is entitled to ask to what extent its support of particular programs in basic science should in due course be linked to the contributions those programs are expected to bring to the public welfare. It is entitled to wonder whether scientists immersed in the excitement of their particular fields, confined by their constraints, and often motivated by the internal structure of their subjects, miss possible developments that would benefit the public. In response, the scientist points to the developments of the past quarter century, during which the technologies emerging from the rapid advance of science in the United States have been most impressive. The skeptics may, of course, reply that perhaps there has been more support for science than needed or justified, no matter how many gains can be cited. These questions are legitimate, and we have no doubt that debate over them will and should continue.

This report is addressed to a limited but crucial segment of the wide spectrum of questions about the federal support of science. It presents

Phase 1 of a study of the peer review system in the National Science Foundation (NSF). By peer review we mean the evaluation of proposals for research in the science disciplines by experts in those disciplines. Recently, there has been both public and congressional concern about the effectiveness and fairness of the peer review system at the NSF. Questions have been raised about whether scientists can be objective as they participate in the distribution within their own community of scarce funds for fundamental research. Could bias or favoritism significantly influence judgments of the quality and promise of research proposals? Could a careful study of peer review as it operates today suggest that some other system would produce more fundamental progress and understanding, greater creativity and innovation in science? The scientific community has been concerned about the public confusion between peer review and the political process by which funds are allocated to the total basic research enterprise. It has also been disturbed by some of the criticisms leveled at the peer review system. Most scientists believe that some form of peer review is the only way of assuring that available funds are being used as efficiently as possible in the development of their disciplines. Scientists see no effective method of deciding the merits of a proposal except by advice from experts in the field. Only such experts are qualified to judge the merit and potential of a proposal. At the same time, the scientific community recognizes that peer review has nothing directly to do with the total allocation of science funds or with the distribution of appropriated funds among the branches of science. It is concerned only with the distribution of funds in rather narrowly defined branches of science.

Because of these public concerns about the peer review process, the National Academy of Sciences (NAS) concluded that a study of the system was in order. The president of the Academy, Philip Handler, asked its Committee on Science and Public Policy (COSPUP) to undertake such a study. It was begun under the leadership of Melvin Calvin of the University of California, Berkeley, past chairman of the COSPUP. The research was done by Stephen Cole, professor of sociology at the State University of New York, Stony Brook; Jonathan R. Cole, professor of sociology at Columbia University; and Leonard Rubin, State University of New York, Stony Brook.

When the COSPUP began considering a study of peer review, Dr. Calvin asked the Coles whether they would be interested in conducting the research, and discussions between the Coles and the COSPUP ensued. As working scientists, the members of the COSPUP had firsthand knowledge of peer review systems, both as applicants for

grants and as reviewers. The Coles had substantial experience in the aspects of the sociology of science pertinent to the study. From these discussions, plans for the study were formulated, and in February 1974, the Coles agreed to do the research. Leonard Rubin, who had just completed his doctorate at the State University of New York, Stony Brook, joined the Coles as a research associate in July 1975. He has conducted most of the qualitative interviews on which a critical part of this report is based.

To study peer review systems, it was essential to have access to information in an agency using such a system. The National Science Foundation and the National Institutes of Health (NIH) were the natural candidates; we chose the NSF. Peer review in the NIH might well be looked at in addition, because there are differences between the peer review processes of the two agencies.

The NSF's general counsel ruled that only a contractor to the NSF could have access to the necessary data. Although we could have sought, and probably received, funds from a private institution in support of this study, the NSF ruling dictated that it could only be carried out by a contract between the NSF and the Academy, with the Coles and Rubin acting as consultants to the COSPUP. To assure free access to essential information, a clause was included in the contract acceptance stating that the NAS would have a right to withdraw if the NSF refused to provide it the data needed to conduct an adequate investigation, or if the NSF in any way hindered the conduct of this investigation. It was not necessary to act upon the terms of that proviso.

A question could be raised concerning whether a study of NSF peer review systems sponsored by the NSF would be biased by reason of that funding support. In our opinion the consultants have been objective throughout the study and have approached the research problem with independence and professional curiosity. Moreover, the COSPUP itself is open to a presumption of bias, for it too is partly sponsored by the NSF. The interest of the committee was in knowing, first, how the system works; second, how well it works; and finally, how it might be made to work more effectively. Therefore, we assured our consultants in advance that they would be free to publish whatever they learned, whether favorable to peer review or not. In fact, a review article by the authors on this research has already appeared in *Scientific American*, October 1977.

What will you find in this study? First, it describes how the NSF decided which proposals were to be funded in 10 programs of their basic science research division for the fiscal year 1975. Second, it gives primary attention to the pivotal role of the program director in the decision-making process. Third, it contains a statistical analysis of the characteristics of applicants, such as age, institution, and citations to previous work. Fourth, it correlates those characteristics with the recipients of grants. In addition, it correlates peer review evaluation with the awarding of grants.

To our minds, the major findings of the study are:

1. *There is a high correlation between review ratings and grants made* (Table 23). The most important determinant of whether or not an applicant receives an NSF grant is the score given his or her proposal by reviewers. Over 90 percent of applicants receiving high scores are awarded grants, and less than 10 percent of applicants receiving relatively low scores are awarded grants. Currently the NSF seems, from this study, to be distributing money largely on the basis of reviewers' evaluations of the quality of the science contained in proposals. Thus, though program directors are not subject to exacting review and may appear to do as they please, they pay very close attention to the review ratings and act in accordance with them.

2. *In the aggregate, there appears not to be a high correlation between grants awarded and the previous scientific achievement ("track record") of the applicant.* On the average, considering all applicants, scientists who have published many papers in the last 10 years and whose papers have been frequently cited by their colleagues have slightly better chances to receive NSF grants than those who have published fewer papers that have received fewer citations. Since one of the stated criteria employed by the NSF in evaluating proposals is the ability of the scientist to conduct the research, it was surprising to us that the correlation between recent productivity and whether or not one receives a grant is as low as it turned out to be. Except for young investigators, the ability to perform research proposed can probably best be judged by consideration of the past record of the proposer; the lasting significance of the proposed research is a different matter, best judged by peer reviews. Perhaps one can explain the surprisingly low correlation between track record and grants awarded along the following lines. Scientists with an excellent track record are very likely to receive a grant. With the exception of young scientists, those with a poor record are unlikely to receive one. These extremes of the population under consideration are dominated numerically by those with track records ranging from fair to very good. Apparently, for this

considerably larger group, reviewers give much greater weight to the merits of the research proposed than to the applicant's previous scientific achievement.

3. *Reviewers residing in major institutions were **not** inclined to treat proposals from scientists at major institutions more favorably.* An analysis of 3,835 pairs of applicants and reviewers indicated that, in fact, reviewers from highly prestigious academic departments were slightly harder on proposals from scientists in highly prestigious departments than were reviewers at less prestigious institutions. And the former were relatively harder on applicants from highly prestigious departments than on those from less prestigious departments than were the latter. In this respect, peer review does not, as has been suggested, serve an "old boy" network in which eminent scientists look after their own interests.

4. *Age had no strong effect on either ratings received or the probability of receiving a grant.* Fifty-four percent of applicants who received their Ph.D. degrees before 1970 received grants; 46 percent did not. Forty-six percent of younger applicants (those receiving their Ph.D. degrees after 1970) received grants; 54 percent did not.

Phase 2 of this study is now in progress. It will evaluate the extent to which the program director affects the awarding of grants by his selection of reviewers. It will also try to determine what types of scientists and institutions make the most effective use of their research grants or contracts. Part of the study will include experiments with "blind" reviews, in which an effort is made to conceal the identity of both the principal investigator and his institution.

The findings in this study are based on the use of several statistical methods. The limitations of those methods are discussed in section 5 and in Appendix B. For the lay reader, the tabular analysis is probably the most meaningful. For example, compare Tables 22 and 23. The proportion of variance explained on ratings by funding history would appear small except possibly for economics. But the tabular analysis in Table 23 shows a definite trend, except for anthropology. Those who received NSF funds in the past 5 years clearly had a better chance of getting higher ratings.

We also want to alert the nonexpert that several linear regression methods, even when applicable, are not fine enough tools to detect the extremes in the range of cases covered in this study. Tables 26, 51, and 52 illustrate the point. Knowing the index, which combines citation and rank of department for a given proposal, does not markedly increase the predictability of whether a proposal will be highly rated. Yet there is considerable difference between the groups at the extremes. In Table 25, 80 percent of those *with a high index of 10* received high ratings, while only 34 percent *with a low index of 2* received high ratings. The proposals falling between the extremes dominate because of their larger numbers, and "on the average," the extremes are not strongly felt. The same can be said of Tables 51 and 52. There is a marked difference in the percentage receiving grants between the highest and lowest ranking. But the value of knowing, say, rank of current department in predicting the funding decisions is not very strong. Again, the group between the extremes dominates on the average.

On behalf of the members of the Committee on Science and Public Policy, I should like to thank Stephen and Jonathan Cole, Leonard Rubin, and, in addition, all those scientists both within the NSF and without who have cooperated in carrying out this study.

I. M. SINGER
Chairman
Committee on Science and Public Policy

Authors' Acknowledgments

We would like to thank the many people who aided us in conducting this study. We owe a special debt to both Melvin Calvin, past chairman, and I. M. Singer, the current chairman of the Committee on Science and Public Policy (COSPUP) of the National Academy of Sciences, for their encouragement and for their suggestions for this research, and we relied heavily upon the exceptional administrative abilities of Robert Green, executive secretary of the Committee. Many members of the COSPUP, past and present, have been very valuable to us in their critical comments made on reports we have presented as the project progressed. Robert K. Merton and Harriet Zuckerman, our colleagues in the Columbia Program in the Sociology of Science, also provided many invaluable critiques of our research as it progressed. Judith M. Tanur has acted and continues to act as a statistical consultant. Jack Kiefer, Donald Ploch, and Burton Singer provided useful methodological advice. Otis Dudley Duncan provided a useful critique of an early draft. Thanks also to Stephen Appold and William Atwood, who did the computer work; to Gloria Lebowitz and Margaret Lardner, who did the typing and transcribing of thousands of pages of field notes and interviews that we collected in the course of conducting this research; and to James Dunne and Clifford C. Hughes, III, who aided in the data collection. Finally, Robert Hume made numerous editorial suggestions that improved the final report.

Throughout this research our liaison with the National Science Foundation was Dr. Jack T. Sanderson, then serving as director of the

Office of Program Planning and Management. We thank Dr. Sanderson for the aid he provided in the data-collection phase of the study. We also thank the many program directors, section heads, and division directors who willingly gave us their time and allowed us to ask them questions, sometimes on several different occasions.

Jonathan Cole thanks the John Simon Guggenheim Foundation for fellowship support during the academic year 1975-1976 and the fellows at the Center for Advanced Study in the Behavioral Sciences for their help during his fellowship year, 1975-1976.

We are, of course, responsible for any errors or misinterpretations of data in this report. We have currently completed only the first phase of our study of the NSF peer review system. A report on the second phase will be ready in about a year.

Contents

Introduction

In 1950 the U.S. Congress established the National Science Foundation (NSF) with the primary purpose of fostering and supporting basic research in the United States. In the 26 years since its founding, the Foundation has grown rapidly both in the amount of money it dispenses for scientific research and in the types of research that it supports. In 1952, the first complete year in which the NSF granted funds, it spent about $3.5 million. Today its budget is approximately $800 million. In 1952 the NSF had 88 full-time staff members. By 1972 the staff had grown to more than 1,000 (Groeneveld *et al.*, 1975, p. 345).

Although originally the NSF was mandated to fund exclusively basic research, in recent years it has been asked to fund some types of applied research in addition to basic research. Thus, the RANN program (Research Applied to National Needs) was established with the aim of supporting research that would be relevant to current national problems. But, while the mission of the NSF has been broadened in recent years, its primary function remains the support of basic scientific research. Indeed, 72 percent of its fiscal year 1976 budget is allocated for the support of basic research. The NSF and the National Institutes of Health (NIH) are today the two primary sources of support for basic research in the United States.

The NSF has gone through several recent internal reorganizations.[1]

[1] To our knowledge these reorganizations had little or no substantive effect on the way in which peer review was conducted.

In part, these reorganizations have been aimed at providing a more rational organization of the growing number of scientific areas that the NSF has been funding. Figure 1 depicts the current organizational structure of the NSF. The National Science Board (NSB) is made up of 24 scientists and laymen who are responsible for setting board policy. Members of the NSB and the director of the Foundation are appointed by the President. The director is responsible for the day-to-day operation of NSF and for carrying out NSB policy.

Currently, the NSF is organized into seven directorates, each headed by an assistant director. The first three directorates—Mathematical and Physical Sciences, and Engineering; Astronomical, Atmospheric, Earth, and Ocean Sciences; Biological, Behavioral, and Social Sciences—fund basic research in the natural and social sciences. In the study reported here we limit our analysis to the decision-making process currently in operation in these three directorates. We are currently studying the decision-making process in Science Education and in RANN. In Table 1 we present the budgetary allocation to the different directorates in 1975-1977.

From its inception the NSF has always received applications for more grants than could be made. Thus some applications have been turned down. We are told by long-time NSF staff members that in years past the great majority of reasonably good proposals were funded. In recent years this situation has changed. The number of competent scientists applying for NSF funds has been increasing, in part because of an increase in the size of many scientific specialties and in part because of

TABLE 1 Budget Program Comparisons, FY 1975-1977 (Millions of Dollars)

Program	Actual FY 1975	Plan FY 1976	Budget Request FY 1977
Mathematical and Physical Sciences and Engineering	180.9	193.4	233.3
Astronomical, Atmospheric, Earth, and Ocean Sciences	184.1	219.3	245.0
Biological, Behavioral, and Social Sciences	104.2	110.4	132.3
Science Education Programs	74.0	64.8	65.0
Research Applied to National Needs	83.6	73.6	64.9
Scientific, Technological, and International Affairs	24.9	22.2	22.0
Program Development and Management	37.9	42.6	43.5
Special Foreign Currency Program	3.6	5.3	6.0
TOTAL	693.2	731.6	812.0

greater difficulty in getting funds for basic scientific research from other federal and private agencies. The Mansfield Amendment, passed in 1972, made it illegal for the Department of Defense to fund any research lacking a clear military application. After this amendment was passed, many mathematicians previously funded by the Department of Defense applied to NSF for funding.

Also, at the same time that numbers of qualified applicants have increased, the cost of doing science has gone up because of inflation. Thus, today the NSF is forced to decline the proposals of many competent scientists. How does the NSF decide which proposals should be funded and which declined? The central element of the procedure used to make these decisions is called the "peer review" system. In general, peer review consists of reviews or evaluations provided by working scientists or "peers"—the peers of the scientists applying for funds.

FORMAL STRUCTURE OF DECISION MAKING

The research directorates have similar administrative structures. The assistant director and staff are responsible for the overall operation of the directorate. The basic organizational units within the directorate are the division, the section, and the program. Each directorate is divided into a number of divisions (representing general disciplinary areas). Some divisions are further divided into sections (representing general subareas within disciplines). Each division or section contains at least one program. For example, the Directorate for Mathematical and Physical Sciences, and Engineering has five divisions: Chemistry, Physics, Mathematical and Computer Sciences, Materials Research, and Engineering. The last three divisions are divided into sections (e.g., Engineering is divided into Engineering Chemistry and Energetics, Engineering Mechanics, and Electrical Sciences and Analysis). The chemistry and physics divisions and the sections within the three other divisions are further divided into programs. (There are 45 programs within this directorate.)

Each structural unit is headed by a person responsible for its operation. The division director is responsible for the functioning of the division; the section head oversees the programs within the section; and the program director has responsibility for the program.

The division directors and the section heads are usually permanent employees of the Foundation. For the most part they are not civil servants but are required to observe many of the same regulations and

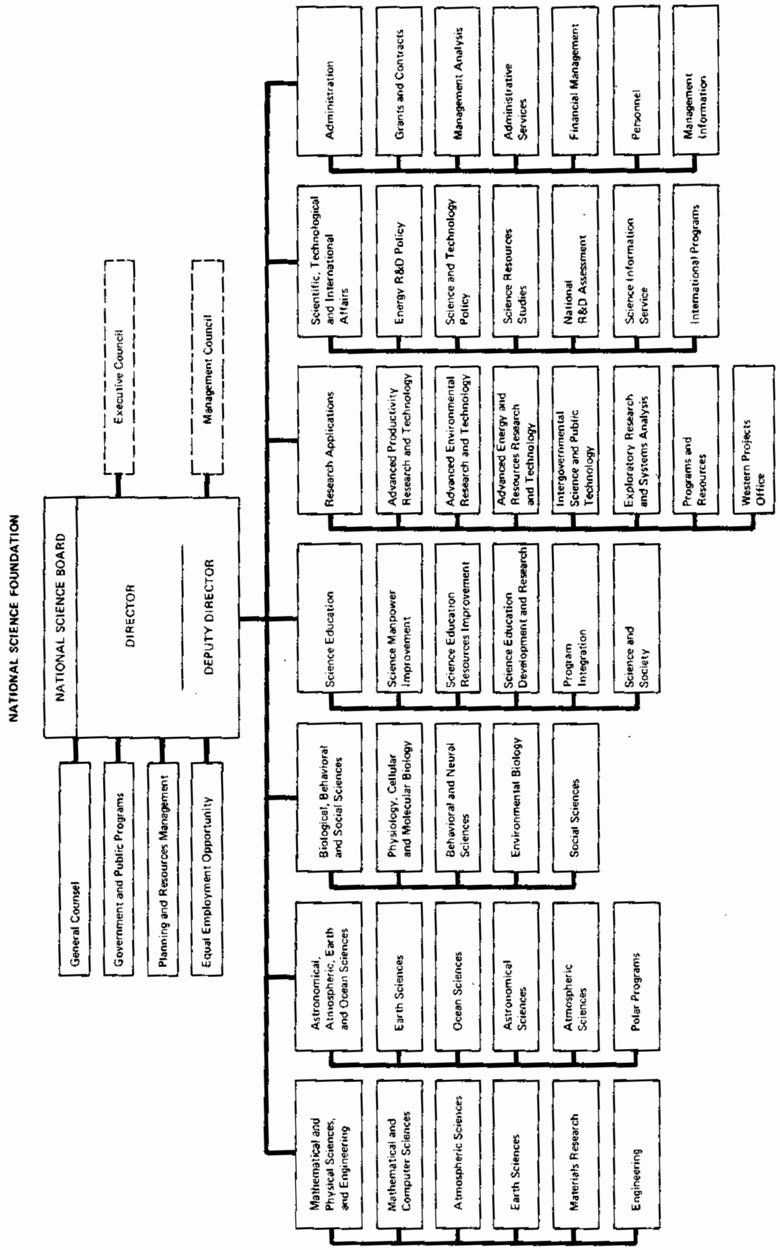

FIGURE 1 Organizational structure of the National Science Foundation in 1975 at the time of Phase 1 of the peer review study.
Source: *Organizational Directory*, NSF, 1974.

seniority rules as are civil servants. The program director, however, can be either a permanent employee or a "rotator." Rotators typically take leaves of absence from their prior affiliations, usually universities, for 1- or 2-year periods to serve as program directors. Currently, about 30 percent of the program directors are rotators.[2]

For the applicant, the most important functionary in the system is the program director. The program is the part of the system with which most members of the scientific community have contact. As we shall see, the program director engages in a number of activities that have a significant influence on funding decisions.

That is the formal structure of the NSF. We now turn to a description of the formal review processes within the Foundation.[3] (See Figure 2 for an illustration of this process.)

Prior to any formal review, of course, a research proposal must be developed and submitted to the Foundation. Before the proposal is submitted, however, the prospective principal investigator may hold preliminary discussions with a program director and may even submit a preliminary proposal. When completed and approved by the inves-
tigator's institution (not necessarily an academic organization, applica-
tions sometimes being received from private research organizations), the proposal is submitted to the Foundation.[4]

Virtually all proposals sent to the NSF go directly to Central Process-
ing and from there are assigned to a division. Some applicants desig-
nate the program they are applying to, in which case the proposal is sent to the designated program office. When applicants make no such designation, Central Processing sends the proposal to a division di-
rector, who, in turn, assigns the proposal to a section. The section head then assigns the proposal to a program. Where there are no sections, of course, the proposal moves directly from the division to the program.

[2]The advantages and disadvantages of each type of program director will be discussed later in the report. Several programs also have associate and assistant program directors, including Mathematics, Chemistry and Biochemistry, Solid-State Sciences, Astronomical Sciences, Biological Sciences, and Social Sciences. Persons in these positions perform different functions. Some, in fact, do the same jobs as program directors and simply share the work of the program with the designated program director. (Associates usually have this role.) Others have as their primary function administrative support. They assist the program director in such tasks as sending out reviews and keeping track of the progress of a proposal in the system. (Assistants usually have this role.)

[3]For a more detailed, comprehensive view of this process, see National Science Foundation, "Peer Review and Proposal Evaluation, Staff Study," *NSF Peer Review Study*, 1975a.

[4]Under the law, individuals without a formal affiliation can submit proposals, but in fact they almost never do.

Notes
1. Principal Investigator (P.I.) conceives research plan.
2. P.I. may hold preliminary discussions with Program Director (P.D.) and/or submit preliminary proposal for evaluation.
3. Final proposal is prepared by P.I., approved by Institution and sent to NSF.
4. Central Processing assigns proposal to a Division.
5. Division Director assigns proposal to a Section.
6. Section Head assigns proposal to a Program.
7a. Program Director chooses reviewers and sends out proposal for independent mail review.
7b. Program Director may choose panel members and send them copies of proposal in preparation for panel meeting.
7c. Program Director may discuss proposal with another agency.
7d. Program Director may discuss proposal with other components of NSF.
7e. Program Director may make site visit (or site visiting team may be appointed and report back to P.D.).
8a. Independent mail reviewers evaluate proposal and return signed, written reviews.
8b. Panel members discuss proposal and indicate rating.
9a. Program Director may decline proposal but suggest some major modification that would make it more acceptable or may suggest that a new proposal may be written.
9b. Program Director may decide to recommend funding and negotiates revised budget with P.I.
10. P.D. recommends funding of revised amount.
11. Section Head reviews recommendation, approves and transmits, or rejects.
12. Division Director reviews recommendation, approves and transmits, or rejects.
13. Assistant Director for Research may review recommendation.
14. If grant is large enough, National Science Board must review.
15. Grant is made to Institution, which disburses funds to P.I.'s project.

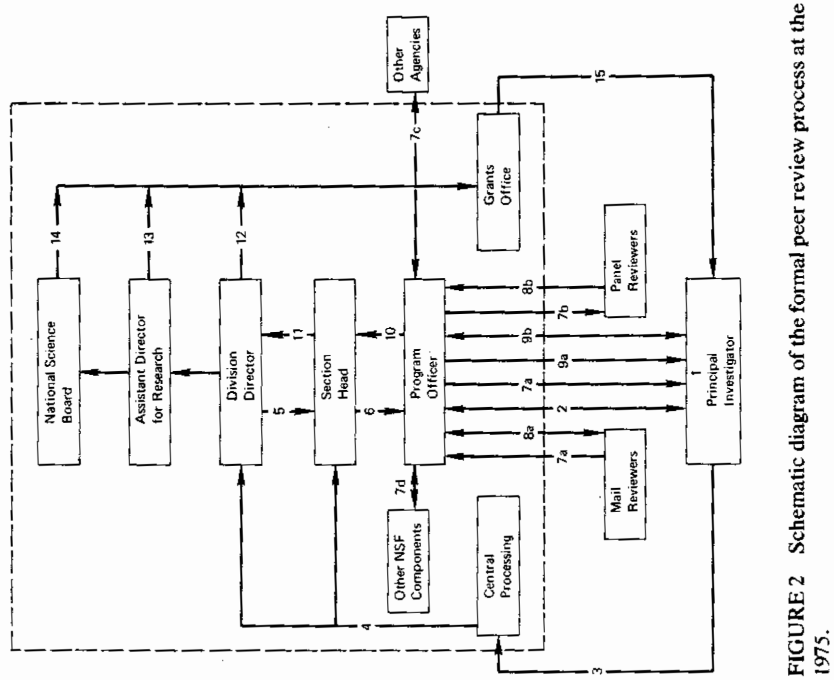

FIGURE 2 Schematic diagram of the formal peer review process at the NSF. Source: "Peer Review and Proposal Evaluation," NSF, 1975.

The review process begins when the proposal is received by the program director. There are two procedures for reviewing basic research proposals in the NSF: *ad hoc* mail review and a combination of panel and mail review.[5] When the *ad hoc* procedure is used, the program director, after examining the proposal, chooses a number of reviewers (about 3-10) and sends them the proposal for independent review. The selection of reviewers—how it is done and how well it works—involves critical decisions. We will discuss this selection process in greater detail in section 2. Along with the proposal, reviewers receive instructions and reviewing forms. Instructions inform them of the criteria they should use in evaluating the proposal. There are 11 stated criteria that are subdivided into four groups: criteria evaluating the principal or named investigator's demonstrated competence; criteria evaluating the content of the proposed science; criteria evaluating the relevance and utility of the proposed research; and criteria evaluating the long-term scientific potential of the research for the United States. (These criteria and the reviewing form, as they appear in the NSF publication, are reproduced in Appendix A.) The reviewing form asks the reviewer for two judgments: an overall adjectival evaluation of the proposal (rating the proposal from excellent to poor) and written comments related to the stated criteria. The program director uses the completed reviews in making his decision.

The stated criteria to be used in decision making include not only an evaluation of the quality of the scientific content of a proposal but also the past performance of the investigator and the ability of his institution to support the research. The formal inclusion of this latter criterion is important, since some people outside the Foundation have intimated that such considerations do not have a legitimate place in the allocation process.

None of the criteria sent to reviewers relates to the matter of the geographic region from which the proposal comes. Yet in the NSF Act approved by Congress and signed into law, Section 2(e) reads: ". . . in exercising the authority and discharging the functions referred to in the foregoing subsection, it shall be one of the objectives of the Foundation to strengthen research and education in the sciences, including independent research by individuals, throughout the United States and *to*

[5]The combination of mail and panel review is used by all programs in the Biological and Behavioral and Social Sciences Directorates and in the divisions of Earth and Ocean Sciences. It is used in Engineering for Research Initiation Grants. The remainder of programs use *ad hoc* mail review.

avoid undue concentration of such research and education."[6] (Emphasis added.) This provision as interpreted by the NSF means that given roughly equal scientific merit, an attempt should be made to see that geographic balance is duly considered in the granting process.

When the combined panel and *ad hoc* reviewing process is used, the program director, after examining the proposal, sends it out to both mail reviewers and panelists. (The size of panels, as well as the number of panelists who receive proposals to review, varies from one program to another.) The panel usually meets in Washington three times a year, and together the panel reviews, mail reviews, and panel discussion provide the basis for the program director's decision.[7]

During the review process the program director may also discuss a proposal with other federal agencies and with people in other parts of the NSF. He may also talk informally with the principal investigator in order to clarify questions about the proposal. In certain cases involving large grants the program director may make a site visit or appoint a team for this purpose.

In the end, after the reviewing has been completed, the responsibility for the decision to fund or not to fund rests with the program director. The peer reviews have an "advisory" status. The program director makes one of five decisions: (1) To fund the project or program. This decision is often followed by further negotiations between the program director and the principal investigator about the budget. These discussions produce a revised budget that the program director recommends for funding. (2) To decline due to lack of funds. (3) To decline but suggest modifications that would make the proposal more acceptable. (4) To decline but recommend that a new proposal be written and submitted. (5) To decline without recommendation.

After the decision is made, it is reviewed by the section head, if there is one, and by the division director. The division director either approves and transmits the decision to the next level or rejects it, thus requiring the program director to reconsider the action. Reconsideration is usually for insufficient documentation; the validity of the decision is not normally questioned.

Recently, a new level of review has been introduced. Each directorate now has an Action Review Board. This review board is

[6]As quoted in National Science Foundation, "Peer Review and Proposal Evaluation, Staff Study," p. 18.

[7]The review panel should not be mistaken for the advisory panel. The latter is a group of 8-10 scientists created on a disciplinary basis. Their function is to advise the Foundation of developments in different scientific areas and to evaluate the performance of individual programs.

chaired by the assistant director or deputy assistant director of the Foundation and is composed of both scientists and nonscientists within the NSF. It meets weekly to review all awards and selected declinations, examining whether projects are consistent with the objectives of the program; whether reviewers were appropriate; whether sufficient consideration was given to their comments; whether NSF grant policies have been followed; and whether the titles of proposed projects appear meaningful to a lay audience as well as to scientists.[8]

In most cases, review by the Action Review Board is the last stage in the review process, but some projects are reviewed by the National Science Board. Such review will take place if a large amount of money is requested ($500,000 or more in a given year or a total of $2 million or more); the commitment is for a period exceeding 5 years; or there are "important policy considerations." Few proposals require review by the NSB.

When the review process has been completed, the investigator and/or institution involved are informed of the Foundation's action. If the proposal is funded, the grant is almost invariably made to the institution with which the investigator is associated, which, in turn, dispenses funds to the principal investigator's project.[9]

Within the past year the NSF has begun to give rejected applicants more information about the content of peer reviews. In June 1975, the NSB established a policy of making available to principal investigators, on request, anonymous verbatim reviews.

COMPARISON OF NSF AND NIH PEER REVIEW

We have compared peer review as it is employed in the National Institutes of Health (NIH), the other major national agency funding basic research, with the way it is employed by the basic research directorates of the National Science Foundation. The NIH uses a review system called "dual review."[10] The NIH does not have program

[8]Letter from H. Guyford Stever, Director, NSF, June 30, 1976.

[9]The following figures provide some sense of the volume of cases dealt with in the review process. In fiscal year 1974 the Research Directorate received about 13,000 proposals. The success rate was approximately 50 percent. On the average, each proposal received five reviews. (SOURCE: Cumulative FY 1974—Statistics of Proposals and Actions.)

[10]For a more detailed description of the peer review process in the NIH see John G. Wirt *et al., R & D Management: Methods Used by Federal Agencies.* See also *Grants Peer Review*, Report to the Director, Phase I, NIH, Washington, D.C., December 1976. This report was released after our Phase 1 research had been completed.

directors who play a determining role in decisions on funding.[11] Rather, the NIH program is divided into approximately 50 study sections headed by executive secretaries. Each of the study sections is a panel of approximately 15 scientists. Each proposal that comes into the NIH is assigned to one of the study sections and is sent out to the members of the study section by an executive secretary. Generally, the executive secretary will ask three members of the study section with particular competence in the area of the proposal to lead the discussion of the proposal. The NIH does not make systematic use of mail reviews.

The study section meets periodically in Washington, D.C., to discuss and vote on the proposals it has received. If a majority of the study section members vote to approve a proposal, each member in secret ballot gives the proposal a priority score from 1 to 5. The executive secretary then averages the scores for each proposal and multiplies the average by 100. This becomes the proposal's total priority score. When the study section has completed its work, the executive secretary arranges all the proposals in order of priority from the top down to those not approved for funding. The NIH staff then identifies proposals that seem to have received inadequate review, have been disapproved by more than two study section members, require funding over $100,000, or are judged to be especially relevant to the Institutes' missions but have not received high priority scores (Wirt *et al.*, 1974, p. 28).

The priority scores and the proposals are then sent to the advisory councils of the concerned Institutes in the NIH. The advisory councils then decide which proposals should be funded, taking into consideration the relevance of the proposed research for the missions of the particular Institutes. But, in general, the priority scores given by the study sections determine the probability of whether or not particular proposals will be funded. Wirt *et al.* (1974, p. 29) point out that over 95 percent of proposals are never discussed at the council meetings. The following are aspects of the NIH system that differentiate it essentially from the NSF system:

1. Proposals are classified according to relevant biomedical problem areas as well as the relevant scientific problem areas.
2. All reviewing is conducted by panels.

[11]The degree to which the NIH executive secretaries influence the decision-making process is an empirical question. Their role is de-emphasized in formal descriptions but may turn out to be more significant.

3. The program manager in the NIH plays an essentially passive role in the decision-making process.
4. Social and medical relevance of research is more significant in the NIH decision-making process.

CRITICISMS OF PEER REVIEW

In recent years several features of the government's decision-making process on the distribution of scientific research funds generally, and the peer review particularly, have come under attack both by government officials and by members of the scientific community.

Perhaps the most far-reaching criticisms of the NSF peer review system were made during a congressional hearing held in July of 1975 by the House Subcommittee on Science, Research and Technology. This hearing was aimed specifically at the NSF peer review system. Approximately 25 witnesses, including former Representative John Conlan of Arizona and Representative Robert Bauman of Maryland, appeared before the Committee to give testimony about various aspects of peer review.

The most important criticism made of the NSF peer review system is that it results in unfair decisions. That is, for example, scientists who are most capable of advancing science are sometimes denied grants and scientists who are doing less significant work are given grants. It was claimed, particularly in testimony given by Congressman Conlan, that the peer review system is essentially an "old boy system":

> I know from studying material provided to me by NSF that this is an "Old Boy System," where program managers rely on trusted friends in the academic community to review their proposals. These friends recommend their friends as reviewers. . . .
>
> Without any effective management control procedures to insure accountability in this kind of system, it is almost inevitable that some program managers may almost unconsciously become advocates for certain scientists and their projects. . . .
>
> It is an incestuous "buddy system" that frequently stifles new ideas and scientific breakthroughs, while carving up the multi-million dollar federal research and education pie in a monopoly game of grantsmanship. [National Science Foundation, *Peer Review*, 1976, p. 40]

It is asserted by critics that this unfair distribution of funds is a result of the extraordinary power that program directors have in deciding who should get the funds. The program director, the critics say, is the agent of an "old boys' club" that gives preferential treatment to the proposals of its members. Eminent scientists make preferential evaluations of the proposals of other eminent scientists to whom they are favorably disposed and deny funds to people who are not part of the "inner circle." The program director is able to lay the ground for this because while his decisions must be reviewed at two levels higher in the organization, in many cases this review is *pro forma*. A recent Office of Management and Budget memorandum (1975) asserted that peer review "produces an unavoidable conflict of interest situation for the scientists who serve as consultants because they determine the allocation of research funds that they also receive" (Office of Management and Budget, 1973, pp. 1-5).

Moreover, say the critics, the reviews received by the program director are only advisory and the program director is free to ignore them. Program directors, it is argued, can predetermine the outcome by selecting reviewers who, they know, will be either hard or lenient on particular proposals. Even if the program director has to make a grant he would prefer not to make, he can effectively stifle the research by reducing the size of the budget. In effect, there is no way of challenging decisions made by the program director.

In order to protect this old boy system, it is claimed, the National Science Foundation cloaks its activities in secrecy, denying congressmen and others access to verbatim reviews and to the names of reviewers of particular proposals. Thus the old boy system is allowed to go on unrestrained, and effective oversight of the NSF by Congress is prevented. It is argued further that the peer review system may stifle innovative research because eminent scientists serving as reviewers may reject ideas that differ from their own.

Other frequent criticisms of the peer review system are:

1. It takes the decision-making power out of the hands of elected officials and their appointees and puts it into the hands of people who are not accountable to the public.[12]

2. It enables the scientific community to use public funds for its own

[12]An OMB memorandum, which has not been published, was distributed to members of the COSPUP at its meeting of June 16-17, 1973. Part V of the document is a summary of criticisms of the peer review system that have appeared over the past 10 years in reports of congressional committees, in articles and letters in *Science*, etc. The document has been reproduced in "National Science Foundation Peer Review Special Oversight Hearings," Subcommittee on Science, Research and Technology, U.S. House of Representatives, July 22-24, 29-31, 1975, pp. 537-544.

purposes, that is, "pure" research, while ignoring the pressing needs of society that might benefit from "applied" research (Office of Management and Budget, 1973, pp. 1-5).

3. It discriminates against scientists working in small science departments at low-prestige universities and colleges.

4. It does not weight adequately the opinions of nonacademic scientists on the merits of proposals. Most mail reviewers and panelists are scientists from prestige universities.

5. It fails to screen out proposals of questionable scientific merit. Senator William Proxmire of Wisconsin has been giving what he calls "golden fleece of the month" awards to projects funded by federal agencies that he believes are of little, if any, merit or utility. Implicit in this criticism is the question of whether the peer review system is sufficient for identifying meritorious research proposals.

REVIEW OF LITERATURE ON PEER REVIEW

Until now there has been very little systematic investigation of how governmental agencies distribute funds for scientific research. The work that has been done can be divided into three categories: (1) general studies of peer review; (2) studies of factors affecting the granting of awards; and (3) studies of outcomes of the review process.

In the first category, perhaps the most thorough investigation of peer review was conducted by the Woolridge Committee in 1965. This study reviewed the peer review process in the NIH and found the decision-making system operating effectively. The Woolridge Committee concluded:

> The opinion of the Committee, based on the extensive investigations of its consultants, is that the large majority of the intramural and extramural research supported by NIH is of high quality. We strongly approve the peer evaluation method of selecting recipients of extramural grants. [Biomedical Science and Its Administration, 1965, p. 3]

The Woolridge Report was essentially a formal description of how peer review in the NIH operates. In 1974, Wirt *et al.* described in a comprehensive report the management of research and development projects in a number of major federal agencies. The report stands as the best source on the formal structure of peer review.

Recently, Thane Gustafson reviewed the literature on peer review and found very little systematic information on how peer review works in the various governmental agencies. However, referring to several "in-house" studies and several other unsystematic studies, Gustafson reviews the criticisms of peer review by Congressmen Bauman and Conlan and concludes that the available information does not warrant any major changes in the peer review system (Gustafson, 1975).

In the second category, a few studies have been done on the effects of characteristics of principal investigators on the probability of receiving grants. Douglass and James (1973) found that between 1966 and 1972, young investigators had a good chance of receiving funds from the NIH. Despite rising declination rates, they found that the proportion of grants going to new investigators remained constant throughout the period.

A study by C. C. Laveck *et al.* (1974) showed that, in the National Institute for Child Health and Development, young people and women have just as good a chance of receiving grants as do older investigators and men. They also found, however, that the young and women received less money on the average than older male investigators.

Small (1974b), in a report to the National Science Foundation entitled, *Report on Citation Counts for National Science Foundation Grant Recipients and Non-Recipients,* found that in most of the fields he investigated, there was a significant difference between the numbers of citations to the work of grant recipients and those to the work of nonrecipients. Grant recipients were for the most part more likely to be highly cited. On the other hand, an internal study conducted by the NSF chemistry section when it was under the direction of M. Kent Wilson reported that the correlation between proposal ratings and numbers of citations to principal investigators was relatively low.

Hensler, in a 1976 report prepared for the Committee on Peer Review, National Science Board, and the Committee on Science and Technology, U.S. House of Representatives, examines the subjective perceptions of NSF peer reviewers and applicants concerning the strengths and weaknesses of the NSF peer review system. In general, these two groups perceived the peer review system as "acceptable" but having some definite weaknesses. Hensler's study is based upon data from a mail survey of 1,068 reviewers of NSF proposals and 2,684 applicants for NSF funds in late 1975 and early 1976. Among other results reported, respondents frequently called for improvements in reviewer-selection procedures (more than a third being in favor of some random selection process), although there was little agreement among them on exactly what those improvements should be. Subjective assessments of the fairness of funding decisions were not related to "academic generation, institutional affiliation or region. . . . About seventy-three percent of the applicants including both grantees and declinees would favor NSF adopting a formal appeals system" (pp. v-vi). Another set of findings suggests that a "majority of reviewers and applicants believe that the NSF peer review process favors proposals from well-known institutions, proposals by older, well established P.I.'s and proposals which are 'in the mainstream' " (p. vii). Hensler acknowledges that her data cannot be used to test the accuracy of these perceptions. The data reported in our study will be useful in testing these perceptions.

A recent article attends to the grant-allocation process at the NSF. Pfeffer *et al.* (1976) investigate the proposition that faculty members on NSF advisory panels give applicants from their institutions a preferred chance of receiving grants. They conclude that this is indeed the case in the four social science disciplines studied: economics, social psychology, sociology, and political science. How this works, however, is not well explicated. Is it a result of particularism or of other unexplained variables, such as greater knowledge of the granting system among faculty members in departments represented on panels?

Groeneveld *et al.* (1975) have studied the social characteristics of advisers to the NSF from 1950 to 1972. Their study leads them to conclude: "Given the high rate of turnover observed, our data suggest . . . that no single group clearly dominates decision-making in advisory positions."

Liebert (1976) has studied the determinants of success in getting grants by examining a subsample of 5,687 from a total of over 40,000 respondents to a 1972-1973 study of faculty members at 259 American senior colleges and universities. The data set did not include information on the size or substance of grants but only on the number received and the source. There were several findings of particular interest. A surprisingly high proportion of the subsample reported having received grants. "Among all scholars in the subsample, 34.5 percent were grant-supported PIs in 1972-73 (21.5 percent in four-year colleges; 44.4 percent in universities)" (p. 666). Liebert was particularly interested in estimating the relative weights of individual productivity and other characteristics of the applicants in predicting the number of self-reported grants. He concluded: In general, we have evidence of a broadly based system for the distribution of research grants that is more competitive with regard to individual productivity criteria than it is biased by field favoritism. There is very little evidence of situational or personal particularism in the all-faculty nationwide data analyzed here" (p. 672).

In the third category, the most systematic study to date of the outcomes of peer review was published by the Rand Corporation in

1974: *Peer Review, Citations, and Biomedical Research Policy: NIH Grants to Medical School Faculty,* by Grace Carter. This report presents data from a study of more than 750 grants made to biomedical scientists working in medical schools and presented two significant findings: first, that the priority score received by a grant on its first evaluation was correlated $r = 0.40$ with the priority score received by the same grant when it was submitted for renewal; second, that the grants that yielded the most highly cited articles were only slightly more likely to have received high-priority scores when they were originally evaluated. Carter interprets these results as being evidence that:

> The later study sections, though composed for the most part of different people, verified the earlier study section's selection of the set of grants that were awarded good enough priority scores to fund. The concept of 'scientific merit' obviously contains enough objective content that different groups of people meeting several years apart will agree that one set of grants is more scientifically meritorious than another set of grants. (Carter, 1974, p. 18)

These moderate correlations, however, could be interpreted as supporting the position that it is very difficult for peer reviewers to predict the extent to which a particular proposal will yield significant research. Carter's study contains analysis only of grants affirmatively acted upon. What we are really interested in is the difference between grants made and grants not made. Are grants being given to scientists who will use the funds most profitably for scientific advance? Although it might be difficult to distinguish grants that will have significant scientific effects from those that will not, it might not be so difficult to distinguish grants that deserve funding from those that do not. Further research is needed on this important question.

Several published studies suggest that peer review is not a precise process. For example, Vivona and DoVan Quy (1973) compared priority scores given to scientific proposals submitted to both the National Institutes of Health and the American Cancer Society. They found a significant correlation between the ratings given, but that it was far from perfect. In some cases one agency gave high priority scores to proposals rated low by the other agency.

Small (1974a), in a report to the NSF entitled *The Characteristics of Frequently Cited Papers in Chemistry,* reports a finding concerning the reliability of peer review judgments. Small found no significant correlation between numbers of citations[13] eventually made to chemistry papers and the evaluations of those papers by journals at the time they were submitted. It has been pointed out frequently that research proposals are more difficult to evaluate than papers submitted for publication. A paper can be judged by something completed, whereas a proposal must be judged by something promised. If predictability based on finished products is weak, we would expect it to be even weaker when based on proposals.

[13]See pages 121-122 for comment on some limitations on use of citation index.

DESCRIPTION OF THE RESEARCH

Review of the literature on peer review suggests that very little is known about how it works in governmental agencies. The research reported on here was directed toward increasing our knowledge in this area. We began by studying how the peer review system works in the parts of the NSF that fund basic research and then moved to studies of other NSF divisions. Data are now being collected for an analysis of how peer review works in the Science Education Directorate. Most of the criticisms of the peer review system within the NSF have been directed toward projects sponsored by the Science Education Directorate, in particular ISIS and MACOS.[14]

In Phase 1 of our research, reported here, our primary purpose is to determine as exactly as we can how peer review works in day-to-day operation of the Foundation. Where does the peer review system in practice diverge from the formal statement of how peer review is supposed to work? Our data are well suited for throwing light on this question and also for pointing up problems with peer review. Problems were revealed in discussions with the people administering the peer review system and by close analysis of the quantitative data. The research is not suited for definitively answering the question of whether the peer review system is an "equitable" one. Although our data allow us to speculate usefully on this question, a more definitive answer awaits the completion of Phase 2 of our research.

Our analysis has led us to believe that probably the most important person in the operation of the peer review system in the NSF is the program director. Therefore, we concentrate throughout the analysis on the tasks performed by the program director and the styles in which they are performed.

The major questions that will be addressed include:

1. What is the role of the program director in determining who gets NSF funds?

[14]ISIS is the acronym for "Individualized Science Instructional System"; MACOS stands for "Man: A Course of Study."

2. Do eminent reviewers give favorable treatment to the proposals of eminent colleagues?

3. To what extent do eminent scientists receive higher ratings on their proposals than noneminent scientists?

4. To what extent do eminent scientists have a better chance of receiving NSF grants than do noneminent scientists?

The findings presented in this report are based on data from three sources.

1. We conducted tape-recorded, in-depth interviews with 70 scientists who have been involved at all points in the peer review system. We concentrated, however, on NSF program directors. We interviewed 35 current and former program directors. We also interviewed members of NSF advisory panels, members of peer review panels, and higher-level NSF officials, including section heads, division heads, and the director and associate director of the Foundation. These interviews ranged in length approximately from 1 to 3 hours. The typewritten transcripts of these interviews constitute a file of more than 2,000 pages of descriptive material on the NSF peer review system. Qualitative interviews such as we have conducted are not suited for testing hypotheses. They are, however, most suitable as descriptive material on how an organization operates and for suggesting hypotheses requiring further research.[15] The purpose of the interviews was not to find out the opinions of specific individuals or how particular program directors or other officials do their jobs, but to find out in a broader way how the organization operates and what some of its problems are.

2. A second source consists of quantitative data collected on 1,200 applicants to the NSF in fiscal year 1975. For each of the 10 NSF programs studied we selected approximately 120 applicants, half of whom were successful and half unsuccessful. The aim of this part of the research was to determine the correlates of getting an NSF grant. We faced a major conceptual decision in selecting a sample of proposals. We could sample a small number of proposals from a large number of programs, or we could be more selective and sample a larger number of proposals from a more limited number of programs. The first alternative would allow us to generalize to all the applications received by the basic science directorates in the Foundation. However, since our preliminary qualitative analysis led us to believe that there were

[15]Since the interviews were conducted with the promise of confidentiality, we will not identify the sources of quotations.

significant differences in the operation of different programs, using all the programs might result in obscuring differences among them. We, therefore, took the second option, selecting 10 different programs and analyzing approximately 120 applications made to each of the programs in fiscal year 1975. A more complete discussion of factors taken into consideration in sampling is presented in Appendix B.

3. Our third source of information was the "jackets"[16] of 200 of the 1,200 applicants. For each of the 10 programs a sample of 20 jackets was selected. We examined the comments by both *ad hoc* reviewers and panel members; the summary and decision of the program director; all correspondence; the proposal; the review of the decision. Notes were taken on these files. Where decisions were ambiguous we reinterviewed program directors with the jackets in hand. We also examined the jackets of about 50 additional cases that our statistical analysis identified as unusual, for example, an extremely eminent scientist whose proposal was declined or a scientist with no past "track record" who received a grant.

This study has particular strengths and weaknesses that should be pointed out. It is, to our knowledge, the first study of its kind to have complete access to confidential files and confidential reviews for both accepted and declined applications. The files and the reviews contained in the files proved to be extremely important in analyzing the significance and meaning of the results of our quantitative analysis.

For lack of time and resources, we limited the study to only parts of the NSF and, within those parts, to only 10 specific programs. We have not yet been able to interview applicants who are dissatisfied with the way they have been treated by the NSF. More importantly, we have no independent measure of the quality of the science proposed in a proposal. In order to answer definitively many of the questions posed by critics of the NSF, we must know whether the science contained in a proposal is of high, medium, or low quality. In Phase 2 of our research we shall obtain an independent measure of quality. Without this indicator many of our results must remain tentative.

[16]"Jackets" refers to the NSF files on particular proposals. For each proposal, the jacket contains comments by *ad hoc* reviewers and, when panels are involved, by panel members; summary and decision of the program director; all correspondence; the proposal; and any review of the decision.

SECTION 1

Program Director Activity Prior to Decision Making

PREPROPOSAL ACTIVITY OF PROGRAM DIRECTORS

Program directors deal with applications that are submitted to the Foundation. Clearly, the population of applicants for NSF grants is not necessarily representative of the population of American scientists. While all studies of the allocation of federal research funds to scientists have concentrated on the procedures employed by the agency in reaching decisions, it is just as important to understand the factors that determine whether or not scientists apply for funds at all. Legally, any person, and certainly any American scientist, has the right to apply to the National Science Foundation for research funds, but many do not.[1]

The number of scientists competing for NSF funds varies, of course, from field to field, depending upon several factors, one of the most important being the availability of other sources of research support. For example, in mathematics apparently very little support is available from other governmental agencies. Therefore, a relatively high proportion of mathematicians apply for funds from the NSF. If we considered all scientists currently employed in Ph.D.-granting institutions in the United States, we would find that a relatively small proportion of those scientists apply for NSF research funds. Factors that determine whether

[1]Useem (1976), in a questionnaire study based upon self-reports, found that 25 percent of anthropologists, 52 percent of economists, 46 percent of political scientists, and 37 percent of psychologists said that in the last 5 years they had not applied for *any* federal funds.

or not an eligible scientist applies for funds are referred to as *self-selection*.

To understand a system distributing limited resources, we must know why some choose to compete and others don't, as well as the procedure used to distinguish among competitors. If the average quality of applicants for funds is high, then deciding among the applicants is very difficult. Correlatively, if there is a great degree of variation in the quality of the science proposed and in the track records of applicants, the task of distinguishing between those who do and those who do not deserve support is somewhat easier. If self-selection mechanisms make decision making more difficult, they also reduce the costs of inefficiency within the decision-making process. If the organization is judged by the quality of the awards that it makes rather than its failure to award meritorious proposals, the average quality of applicants will in large measure determine the quality of the job the agency is doing.

We have collected limited, and thus incomplete, data on the self-selection of applicants to the NSF. These data require further analysis and study. On the basis of preliminary examination, these data clearly indicate that, on average, applicants for NSF funds have more impressive track records as scientists than either American scientists in general or scientists at Ph.D.-granting departments rated in the 1971 American Council on Education (ACE) study of graduate institutions (Roose and Andersen, 1971). In evaluating the results presented throughout this report, we should keep in mind the probability that we are dealing with applicants who are not representative of American science as a whole, but who are more representative of productive research-oriented scientists. We examine next the extent to which the activities of NSF program directors influence the types of proposals they receive.

Prior to the actual review process, program directors engage in a variety of activities that significantly affect what is ultimately funded. These activities can affect both the areas that may or may not be funded by a program and the form of the proposals that are submitted to the program.

According to the NSF, the research proposed for funding by the scientific community is more a function of that community's independent assessment of the direction that research should take than of what NSF staff deems to be significant. In this view, the Foundation is quite passive; it elicits scientific judgment and acts on the basis of that judgment by providing material support for research. According to this view, the program director does not have to play an active role in seeking particular proposals. This process of "notification" by the scientific community has been referred to as "proposal pressure." The content of the program, then, is largely determined by these outside judgments.

This view sees the program director functioning to maintain a network that will facilitate this information exchange. One program director described his role in this regard:

> We do not see our role as pushing the community into whatever we perceive as important to us. We do see our role as trying to identify in the scientific community what people who are particularly capable think is important.

Other program directors expressed a similar point of view; they saw their role as a reactive one. However, some program directors described themselves as being much more active in shaping the direction of research in their programs. They did not agree that proposal pressure was always a good indicator of the directions programs should take. Some of them pointed out that scientists respond to fads and that sometimes the most faddish topics are not the most important ones.

One program director talked about his orientation in the following way:

> I have, contrary to the usual practice in this division, publicized rather widely areas where I thought we were making some advances, or where we were not getting good ideas, or where we were just getting replications where we didn't need them. I have gone around suggesting that people develop proposals in certain areas and saying that if we could show development in these areas, we could probably increase budgets to sustain work in these areas.

This statement suggests that on the basis of his own judgment and the opinion of members of the relevant scientific community, the program director can and sometimes does take the initiative to stimulate certain lines of inquiry. Thus, he not only assesses the state of affairs in his field but also can try to facilitate or impede certain kinds of research.

In short, program directors adopt differing styles in stimulating research. They can be influential in determining who applies to the program. Program directors who are active in shaping the substance of their programs may, advertently or inadvertently, cause some people to decide not to apply for NSF funds. This is where self-selection and social selection (or institutional selection) merge. If it becomes widely known that a particular program director favors one type of work over others, it should not be surprising when he receives few proposals representing an out-of-favor work style. The extent to which potential applicants consider the preferences of program directors is worth further investigation.

Program directors can influence the types of proposals submitted to the NSF through contact with prospective applicants. All program directors interviewed acknowledged that they had had some contact with prospective principal investigators prior to submission of formal research proposals. This contact ranges from telephone calls to program directors to inquire whether there is interest in particular areas to submittal of draft proposals as means of preliminary exchanges of ideas between prospective principal investigators and program directors.

Attitudes toward contact with applicants prior to formal proposal submission vary among program directors. Some actively encourage such contact and, in fact, are quite specific about what they would like to see. One person said the following:

> There has been an elaborate process of critiquing preproposals and in trying to send out clear signals as to what is a compatible (with the program) proposal and what is not. I think it has improved the quality of the proposals we get.

Others encourage contact but insist that program directors should not try to encourage applications in specific areas. One person said this:

> Probably 50 percent of the people talk to me for either real or imagined reasons before submitting. Often people will ask questions about how the system works or are we interested in their particular thing. At times it takes me a long time to convince them that I am really interested in the best science that money can afford and that I don't have a particular shopping list.

Another talked about how his involvement in this activity was mostly with young scientists. He said:

> Actually, what I'm doing is giving them a course in grantsmanship, the art of writing a proposal. A young person can have a delightful idea, but if he presents it in a crummy way, he's going to get zapped. What I want to do is make him competitive enough so that people will not be turned off by the way he presents the thing.

Other views about preproposal contact are seen in the statements by the following program directors. One said:

> I don't like to do it because it does put you in some sort of a position to make some kind of judgment about it beforehand.

Another put it this way:

> We avoid that [preproposal consultation] like the plague. We refuse to give them ideas about how it ought to be written. Otherwise, that would put us in a bad way if we helped write it.

The attitudes and actions differ, but the possibility of preproposal contact with the program director—officially neither proscribed nor prescribed—exists for all. These negotiations can either "cool out" applicants so that they do not even apply or "heat them up" so that they do. To what extent do prior negotiations make explicit to scientists the standards of review? What types of scientists use these informal contacts? How are they used? Is prior contact useful in reducing the number of substandard proposals submitted? Although we have no direct evidence on this, scientists who have been funded by a program for a long time probably are more likely to discuss their renewal applications with program directors prior to formal submission than are new applicants. Further work on this aspect of the process is required to provide answers to these questions.

In sum, there is considerable variation in the extent to which program directors attempt to influence the forms that their programs will take. Regardless of how much initiative the program director takes, it is important that the opportunity for such initiative is permitted by his position in the structure of the organization.

PREREVIEW EVALUATION BY PROGRAM DIRECTORS

When a proposal is sent to the program, the program director decides whether or not it has been correctly assigned to his program. He can accept the proposal or he can attempt to have it transferred to a more appropriate program. If the proposal cannot be reassigned, it remains where it was originally sent, but it is at some disadvantage, since the program director may see its topic as marginal to his field.

After the proposal has been accepted for review by a program, the process of evaluation begins. The first step in this process usually involves some initial screening of the proposal by the program director. Again, program directors vary greatly in the degree to which they get involved in the initial review process.

Some directors read every proposal and make some kind of preliminary evaluation. The following illustrates this:

> I would read every one of them. I have sometimes a long page written but usually a short page written on each one before the panel meeting. I would also have a number written down in the corner which nobody else would see—it was my own evaluation of it.

But many program directors give proposals only a cursory scanning before sending them out for review. One described it in the following way:

> I don't read it in detail, usually. I pick a list of reviewers and send a form letter along with it.

These differences can be attributed to a number of factors. As we have noted, program directors have different styles; they have a good deal of discretion in designing review procedures. Another factor is the volume of proposals, which affects the amount of time a program director has for initial reviews.

The presence and size of ancillary staff (assistant and associate program directors) assisting with administrative details, for example, logging reviewers used and completed reviews, may influence the extent to which the program director can become involved. Such support allows the program director to spend more time in reading and evaluating proposals.

In addition to preliminary evaluation, other kinds of assessments are made at this stage. Some program directors spoke of a sorting process, which one described as follows:

> I have certain subdisciplines within . . . What I will do is take the proposals and pile them up according to their subdiscipline—the ones that I want to compare with one another, the ones that can be compared with one another by experts outside.

The program director quoted above talked about decisions made on the basis of the substance of proposals. Others talked about making initial distinctions on the basis of the characteristics of applicants. One said:

> We do the screening and see if it is a very well-established scientist who is already receiving a large research program but going for additional support, is it a young investigator just getting started, is it somebody so far out of it that this proposal is not even a proposal?

This preliminary work leads to selecting both who will review and how many reviewers will be chosen. Regardless of any evaluation, all proposals do go out for review. No cases were reported in which no reviews were solicited. Since time and work pressure is great, decisions made at this point are intended to make the review process as efficient as it can be. One program director described such a decision:

> We might decide that a proposal that we consider is a poor proposal (and these are very few and far between) might only be sent to three or four people for review. I think the decision is made then simply on the basis that it takes a long time to review a proposal, and we don't want to waste people's time by requiring them to review poor proposals.

Thus, this work sets the stage for perhaps the most significant aspect of the review process—the selection of scientists to review proposals.

SELECTION OF PANEL MEMBERS

Some NSF programs use review panels in addition to *ad hoc* mail reviews. Program directors play a central role in selecting panel members and in administering panel reviews. Panel members are selected by program directors after consulting with section heads or division directors, or both. According to our interviews, the program director's choice of panelists is rarely overruled by his supervisors. It is interesting to note that, at the NIH, where all proposals are reviewed by panels or "study groups," panelists are selected by the executive secretaries. Thus, while the role of the NIH executive secretary is assumed to be less influential than that of the NSF program director, they have the same discretion in selecting panel reviewers.

Our interviews with program directors and section heads identified seven considerations that affect the recruitment of panelists in varying degree. These are:

1. A balance of substantive interests on the panel. Proposals in different substantive areas require an appropriate range of expertise among panelists.
2. Broad general competence. Although their specific expertise is needed, it is also desirable to have panelists who can evaluate the broader implications of proposed research.
3. The background of the program director. He will try to select people who are knowledgeable in areas in which he is not.
4. Geographic distribution. An attempt is made to ensure that all regions of the country are represented.
5. Age. The panel should include both younger and older scientists.
6. Ability of panelists to handle a heavy reviewing load. Panel review is a demanding process.
7. Ability of panelists to work together. They must be compatible and mutually responsive.

What sources of information do program directors use in selecting panelists? In some cases the program director has no direct knowledge of people in certain areas. He then has to rely on his professional networks. As one put it:

> In areas where I was not particularly familiar, I would go through two steps. One, I would talk to people within the Foundation who had expertise and knowledge and solicit names from them. Secondly, I would go to professional societies and solicit suggestions from them.

Personal experience probably remains the most important factor in selecting panelists. Many program directors talked of obtaining recommendations from present panelists about possible replacements. One described it this way:

> You consult your panel alumni and your present panelists. They throw into the hopper suggestions. This could be viewed negatively as a self-perpetuating dynasty. I view it as input from people whom I respect.

It is possible to view the idiosyncratic and personalistic ways in which most program directors select panel members as being part of an "old boy system." Concluding that panelists are frequently selected through an old boy system, however, does not tell us whether the method of selection influences the decision made. Panelists selected in a particularistic way might make very equitable decisions, and panelists selected in a universalistic way might make very inequitable decisions. We shall address this question again later on in the report.

After panels have been chosen, program directors assign proposals to panelists. This varies from program to program. In some cases, panelists receive all the proposals for a particular session and are free to review as many as they want (the expectation being that they will certainly review those that fall within their areas of expertise). In other cases, panelists are assigned specific proposals to review. Not all panelists are required to write reviews.

The program director has considerable administrative responsibilities in this reviewing system. He must initiate and process two types of reviews—panel and mail. After the panelists return their ratings to the program director, he records them on a tally sheet that is presented to the panelists at a session in Washington. During the discussion of a specific proposal, the program director may present some of the *ad hoc* reviewers' comments to the panelists.

SELECTION OF MAIL REVIEWERS

As we noted, one frequent criticism of the NSF peer review system is that the program director has the opportunity to select a biased set of reviewers, which will ensure a given outcome. Former Congressman John Conlan maintained that:

> It is common knowledge in the science community that NSF program managers can get whatever answer they want out of the peer review system to justify

their decision to reject or fund a particular proposal. . . . Since program managers soon learn, like college students, which professor is good for an easy "A" and which can be counted on for an almost certain "C" or "D," it's no trick to rig the system. [Peer Review hearings]

There are two dimensions to Mr. Conlan's claim: First, that the system is structured to permit program directors to "fix" reviews; and second, that program directors frequently take advantage of this opportunity.

Every program director interviewed was asked to react to this claim. Most agreed that Conlan's first contention is valid. It is *possible* for program directors to select reviewers who will give particular types of reviews. One program director spoke for many:

They [program directors] certainly can manipulate reviews. Again, it's easy if you know anything at all about your reviewers. Like in sending proposals to three reviewers who rate anything I send them as excellent. The same way that I can find three cranks who rate anything as poor. In principle I could do that.

However, not surprisingly, program directors almost uniformly disagree with Conlan's second contention that they actually take advantage of the possibility. The person who was just quoted said in this regard:

The statement that it can happen is a very different statement from the statement that it does happen. I am not aware of it ever happening. I think you are more likely to find evidence of incompetence among program staff than evidence of intentional manipulation.

Other program directors also insisted on making this distinction between what could happen and what actually goes on. Some talked of hearing about such abuses, but they maintained that the extent of such activity was extremely limited.

Given the possibility of transgressions, why are "fixed reviews" so unlikely? First, many of the program directors maintained that predicting outcomes is not as easy as it might appear to be. As one person put it:

Out of the 100 or so people that I use, I can think of 1 who is somewhat predictable. But I've seen people send in three fairly tough reviews and then come back with a relatively easy one. It's very difficult to tell.

Second, the volume of reviewing done in a program can make it extremely difficult to know in advance how reviewers would respond:

We process 300 proposals a year—each one gets 5 reviews and we try not to use the same guy more than once or twice a year. It couldn't be possible to fix reviews with the large number of proposals that we have.

Third, most program directors claim they have nothing to gain by manipulating the process. One said:

I worry about power only when the person wielding the power has something to gain from it, and I can't for the life of me see what he can gain except the knowledge that he has listened to the people who are in the field and has managed to have the field going in the right direction.

Fourth, the program director is accountable to the larger scientific community, and such accountability prohibits acting in a self-interested manner. As one person put it:

How can a man who serves as a program director support abuses and continue to be a practicing, reliable, honorable member of the community? He can't do it. In a sense there is a guarantee, because if a person did send out proposals to be reviewed by "cronies" everyone would soon know about it and they wouldn't want him as a program director.

In sum, program directors believe that the possibility of manipulating reviewer selection does exist, and some program directors believe that manipulation goes on to a limited extent. Despite these mitigating statements by program directors, however, the question of manipulation of reviewer selection remains serious. At this point, we have no way of determining precisely the extent of such abuse. Although program directors claim that there is not, in fact, a significant amount of "review fixing" and are able to state reasons for its absence, they are clearly interested parties in the dispute. We cannot expect them to admit to widespread bias. Since bias in selecting reviewers is a crucial issue, we have investigated it further. In the next section we present quantitative data on how selection of particular types of reviewers may or may not affect the outcomes of the decision-making process. But even these data will allow us to go only so far. In Phase 2 of our research, we are conducting a study that will attempt to answer this question. We plan to send a set of proposals already processed by the Foundation out to a group of reviewers selected by knowledgeable scientists not connected with the NSF. We shall then compare the ratings given by these two groups. If the correlation between the two sets of ratings is high, this will constitute some evidence that any NSF bias in the selection of reviewers has little significant impact on the outcome of the decision-making process.

Putting aside the question of "review fixing," what sources do program directors use to obtain knowledgeable reviewers? Among the variety of sources are: personal knowledge of the field, professional contacts in the field, references used in the applicant's proposal, files in the NSF offices, journal articles, and proceedings of professional meetings.

I think the selection of mail reviewers is mostly from my personal knowledge of the field of ______________. I've given seminars at probably 150 universities during the course of the 15 years that I've been teaching and I know people. I read the journals and we have boxes filled with lists of reviewers that have been used in the past.

And:

I have several kinds of lists of reviewers; previously used reviewers. I have journals and proceedings of professional meetings. I could also look at the references in the proposal. I also have a list of the faculty members at all the universities.

Locating sources of potential reviewers is apparently somewhat less difficult than deciding who to ask to review a proposal. A large number of factors influence the choice of who and how many should review. Most program directors tried to get a mix of general and specific reviews on a proposal. One person put it this way:

Basically, I try to reach guys who are highly qualified in the field. However, I try to pick one of the group who is not as specialized in that area. He is familiar with it, but he can stand back and look at the field from a slightly different viewpoint.

The need for care in reviewer selection is especially great when work is being proposed in a somewhat controversial area. Many program directors talked about this type of case and the ways in which they handle reviewer selection. One said:

We try to send it to three types of people in these cases. You send it to the sympathetic ones, knowing their bias. You send it to some known negative critic to see if he can distill out the substance. And in between you have to rely on people who are more general, generally competent people who don't fall into a camp, but who can give you a more or less objective view on the proposal's strong points.

Another program director spoke of his more general approach to the problem:

You have to have a good knowledge of the subfields—who are working in them, what are their conflicts. You have to then have a calibration on the

reviewer. You ask, "What are the conflicts—why this reviewer might or might not give you a good one."

Sometimes the research areas are so small that there is no way to get balance among reviewers. One program director described such a situation:

We had a difficult time for a while because we didn't have people outside that "school" who were qualified to review proposals. I think a lot of proposals were funded in that area and we weren't getting good independent critical judgments.

The kind of work that the principal investigator is doing also affects reviewer selection. A physicist program director commented:

We try to make a reasonable balance between experimental people and theoretical people in the field if there is a significant theoretical component of the proposal. If it is a very large proposal with the operation in a lab, you try to select some people who have had experience managing a lab in addition to the straight physics.

It's not very often that I will ask the average experimentalist to review a theoretical work, because normally that's not a good idea. However, a very good experimentalist, with theoretical overtone, will review a theoretical proposal. But more likely, I will take an experimental proposal and ask a theoretician to review it because a good theoretician is always looking at experimental results—that's where he starts from, that's where he leaves off.

Some program directors try to balance industrial and university people, in fields in which this mix pertains. One director said in this regard:

We try to get reviews from both industry and from the university. We want to get advice from practitioners, and depending on the subject matter, we can get very perceptive damning reviews from some of the industrial people.

Another set of considerations is possible connections between applicants and the reviewers. One program director said:

I check their bibliography to see whether he has any past connection or collaborative effort with the man or if he was his thesis adviser; so I avoid people who are colleagues.

The relative eminence of reviewer and proposer is also considered. A number of program directors spoke of this. One said:

When I get a proposal from a great man. I would use at least two other great men in reviews. The problem with using young reviewers versus established is that young reviewers are apt to give innocuous reviews in these cases. A more experienced reviewer is inclined to say what he thinks one way or another.

A further consideration in selecting reviewers is their "track record." One program director said:

> There is one man I stopped using because anything I sent him he said was awful and he didn't give me any information.

Another expressed a similar view:

> There is one type of reviewer that I tend to eliminate—the type that always gives a negative review and a very low rating. This man will give it a low one, no matter what it is.

Finally, how reviewers fulfill their obligation to return reviews affects their selection. One program director said:

> Over the years we have built up this list of adequate reviewers on the basis of the number of times they returned it when you asked them. If you send them one or two proposals three times a year and you get back one a year, you don't send any more.

In sum, considerations relative to selection of reviewers are numerous. Every program director stated that selection of reviewers is extremely difficult and, also, that it is at the heart of the process. Directors must have considerable scientific and administrative expertise to make the kinds of decisions that will lead to the best possible reviews for proposals. It should also be pointed out that, although the program directors as a group are aware of the many factors that must be considered in selecting unbiased reviewers, errors are undoubtedly made.

SECTION 2

Relation between Reviewer and Applicant Characteristics as an Influence on Ratings

Critics claim that the program director can predetermine the outcome of the peer review process by sending proposals to scientists whose evaluations of the proposals are predictable. This might be termed the "old boy hypothesis," which presumes that the proposals of eminent scientists who are members of the "old boy network" are sent to other eminent scientists who give the "old boys" favorable evaluations. Equally important, the proposals of noneminent scientists, who are not part of this network, are sent to scientists who will give them lower evaluations than they deserve. Although we have no evidence one way or the other that the program directors select reviewers with a certain outcome in mind, we can, by looking at the outcomes, see whether the data support the old boy hypothesis. Do eminent reviewers actually rate the proposals of eminent colleagues more favorably than other reviewers?

An immediate problem in testing the old boy hypothesis is the absence of conceptual clarity in the charge. The charge is that research money is allocated unfairly, but the attribution of this unfairness to "old boyism" is somewhat confusing. What is referred to by the label old boy network? There are at least three possibilities. It could refer to scientists with a common view of their fields who will favorably appraise work only by others with similar views. It could refer to social networks of friendship—made up of scientists who know each other, "grew up" together, attended the same schools, tend to fraternize, are of the same sex, and favor each other's research proposals. It could refer to social positions; that is, those scientists who have achieved eminence tend to favor the proposals of others who are similarly situated in the hierarchy of science even if they have no personal contact with them. Critics of the peer review system never specify clearly the forms of old boyism that undermine the peer review system. The data we have collected allow us to examine the claim that persons of similar rank, intellectual background, and repute favor each others' proposals. We do not have data to examine the other forms of old boyism connected with friendship patterns.

Analysis of the 1,200 applicants for NSF funds provides some data relevant to the old boy hypothesis. For each of the 10 programs we studied in detail, we have data on the characteristics of principal investigators and reviewers and on numerical ratings given. We have begun the analysis of these data by examining the rankings of the applicants' and reviewers' current departments. The data on the 10 programs are presented in Tables 2-11. They enable us to answer five questions:

TABLE 2 Rank of Department of Reviewers by Rank of Department of Applicants: Algebra

Rank of Department	Applicants, %	Reviewers, %
Top 15	17	43
Other ranked	42	25
Unranked and nonacademic	42	32
TOTAL	101	100

	Rank of Department of Reviewers, %			
Rank of Applicants	Top 15	Other Ranked	Unranked and Nonacademic	Total
Top 15	60	15	25	100
Other ranked	45	28	27	100
Unranked and nonacademic	34	25	40	99

	Rank of Department of Reviewers, Mean Ratings Given			
Rank of Applicants	Top 15	Other Ranked	Unranked and Nonacademic	Total
Top 15	1.98	1.31	1.69	1.81
Other ranked	1.75	2.14	2.00	1.92
Unranked and nonacademic	2.13	2.70	2.21	2.31
TOTAL	1.93	2.29	2.07	2.06

TABLE 3 Rank of Department of Reviewers by Rank of Department of Applicants: Anthropology

Rank of Department	Applicants, %	Reviewers, %
Top 10	12	37
Other ranked	27	15
Unranked and nonacademic	62	49
TOTAL	101	101

	Rank of Department of Reviewers, %			
Rank of Applicants	Top 10	Other Ranked	Unranked and Nonacademic	Total
Top 10	50	7	43	100
Other ranked	30	19	52	101
Unranked and nonacademic	37	14	49	100

	Rank of Department of Reviewers, Mean Ratings Given			
Rank of Applicants	Top 10	Other Ranked	Unranked and Nonacademic	Total
Top 10	2.68	–[a]	2.67	2.55
Other ranked	2.38	2.55	2.75	2.61
Unranked and nonacademic	2.74	1.95	2.34	2.44
TOTAL	2.66	2.09	2.50	2.50

[a]Not enough cases.

1. What is the distribution of applicants and reviewers among different-ranked departments? Are reviewers more or less likely to be drawn from top-ranked departments than are applicants?

2. Are proposals from applicants currently employed in top-ranked departments more likely to be reviewed by reviewers from top-ranked departments than are proposals from applicants currently employed at less prestigious institutions?

3. Are applicants from top-ranked departments more likely to receive favorable ratings than are applicants from lower-ranked departments?

4. Are reviewers from top-ranked departments more or less lenient in their ratings than reviewers not from top-ranked departments?

5. Are reviewers from top-ranked departments more likely to favor proposals from top-ranked departments than are reviewers from lower-ranked departments?

In order to respond to these questions, we shall examine in detail the results for 1 of the 10 programs, algebra (see Table 2). The top of the

TABLE 4 Rank of Department of Reviewers by Rank of Department of Applicants: Biochemistry

Rank of Department	Applicants, %	Reviewers, %
Top 15	12	21
16-32	14	18
Other ranked	28	16
Unranked and nonacademic	45	45
TOTAL	99	100

	Rank of Department of Reviewers, %				
Rank of Applicants	Top 15	16-32	Other Ranked	Unranked and Nonacademic	Total
Top 15	17	17	17	49	100
16-32	13	22	15	50	100
Other ranked	22	20	16	42	100
Unranked and nonacademic	24	17	15	45	101

	Rank of Department of Reviewers, Mean Ratings Given				
Rank of Applicants	Top 15	16-32	Other Ranked	Unranked and Nonacademic	Total
Top 15	1.56	2.06	2.39	1.92	1.96
16-32	2.25	1.79	2.39	2.27	2.18
Other ranked	2.81	2.37	2.58	2.65	2.62
Unranked and nonacademic	2.79	2.91	2.66	2.79	2.79
TOTAL	2.62	2.47	2.56	2.56	2.56

table shows the distribution of both applicants and reviewers among different types of departments. In algebra less than one-fifth of the applicants are employed in the 15 top-ranked mathematics departments.[1] A substantial portion (42 percent) of the applicants are currently employed either in departments that are unranked or in nonacademic jobs. The reviewers are far more likely to be in the top-ranked departments than are the applicants. Forty-three percent of the reviewers are currently employed in the top 15 departments.

Do these data by themselves tell us anything about the equity of the review process? The fact that reviewers are drawn heavily from prestigious departments tells us nothing about it. Presumably, there is some concentration of talented algebraists in the most prestigious

[1]Throughout the analysis we have been forced to use the general disciplinary ratings available in Roose and Andersen (1971). There are no departmental ratings for specialties like algebra. We must therefore assume that high-ranking mathematics departments are in general the most desirable places for algebraists to work.

TABLE 5 Rank of Department of Reviewers by Rank of Department of Applicants: Chemical Dynamics

Rank of Department	Applicants, %	Reviewers, %
Top 15	13	27
16-40	25	16
41-70	30	24
Unranked and nonacademic	32	33
TOTAL	100	100

	Rank of Department of Reviewers, %				
Rank of Applicants	Top 15	16-40	41-70	Unranked and Nonacademic	Total
Top 15	27	24	21	29	100
16-40	30	16	22	33	100
41-70	25	12	25	38	100
Unranked and nonacademic	30	14	26	30	100

	Rank of Department of Reviewers, Mean Ratings Given				
Rank of Applicants	Top 15	16-40	41-70	Unranked and Nonacademic	Total
Top 15	1.70	2.05	1.82	2.19	1.94
16-40	1.84	2.21	2.39	2.44	2.23
41-70	2.14	2.59	2.03	2.12	2.17
Unranked and nonacademic	2.16	2.64	2.39	2.59	2.42
TOTAL	2.01	2.41	2.22	2.34	2.24

departments. Many studies have shown that mean faculty prestige and productivity are highly correlated with departmental prestige (Cole and Zuckerman, 1976). Program directors seek reviews, of course, from the best people in the field, and these people tend to be concentrated in the top-ranked departments.

In the second part of Table 2, we report the distribution of ranks of departments of reviewers of proposals from different-ranked departments. In algebra the program director was more likely to assign proposals from applicants in prestigious departments to reviewers in prestigious departments. While 60 percent of the reviewers of proposals from "top-15" applicants were in top-15 departments, only 34 percent of the reviewers of proposals from unranked and nonacademic applicants are located in top-15 departments.[2]

[2]Assuming that old boyism is indeed at work, its influence on outcomes is in part dependent upon the extent to which proposals of eminent applicants are disproportionately reviewed by eminent reviewers. The stronger this relationship, the greater the *potential* influence of old boyism.

TABLE 6 Rank of Department of Reviewers by Rank of Department of Applicants: Ecology

Rank of Department	Applicants, %	Reviewers, %
Top 18	37	32
19-50	23	16
Unranked and nonacademic	40	52
TOTAL	100	100

Rank of Applicants	Rank of Department of Reviewers, %			
	Top 18	19-50	Unranked and Nonacademic	Total
Top 18	33	16	51	100
19-50	38	14	48	99
Unranked and nonacademic	30	17	53	100

Rank of Applicants	Rank of Department of Reviewers, Mean Ratings Given			
	Top 18	19-50	Unranked and Nonacademic	Total
Top 18	2.58	2.22	1.87	2.16
19-50	2.18	2.35	2.16	2.19
Unranked and nonacademic	2.63	3.05	2.44	2.59
TOTAL	2.50	2.59	2.18	2.34

At first glance, these data may be seen as offering some support for the old boy assumption. Such a conclusion would be erroneous. We do not yet know whether reviewers from top-15 departments are likely to favor applicants from top-15 departments.

The necessary data are presented in the bottom part of Table 2 (low rating = favorable, and high rating = unfavorable). The total column shows that applicants from high-ranked departments are indeed more likely to receive favorable ratings than applicants from lower-ranked departments (comparing the 1.81 with the 2.31). This fits our expectation, since we know that the most productive scientists tend to be concentrated in the prestigious departments. The column totals show that top-15 reviewers are in general slightly more likely to give favorable (low) ratings than are mathematicians employed at less prestigious institutions (comparing the 1.93 with the 2.29 and the 2.07).

The crucial question, however, is whether top-15 reviewers rate proposals from top-15 applicants more favorably than do reviewers from other institutions. The answer to this question for algebra is an unambiguous "no." Top-15 reviewers are tougher on proposals from top-15 applicants than are other reviewers. The mean review given by reviewers in top-15 departments to proposals submitted by applicants in top-15 departments is 1.98. This is a less favorable score than the mean review given by reviewers in lower-ranked departments. The information from Table 2 is summarized in the first row of Table 12. The first statistical test performed is a comparison of the mean rating of applicants from the top group of departments with the general mean rating. As we see in algebra, this is statistically significant at the 0.025 level. The figures in this part of the table simply tell us whether applicants from top-ranked departments get on the average more favorable ratings than do applicants in other departments. They do, except in anthropology and meteorology. In the other eight programs, the difference is statistically significant at the 0.05 level or less.

The next section of Table 12 indicates whether the reviewers from top-ranked departments are more likely to be lenient or tough than are reviewers from other departments. Since the mean rating given by top-ranked reviewers in algebra is lower than the mean rating given by all reviewers and the difference is statistically significant, we can

TABLE 7 Rank of Department of Reviewers by Rank of Department of Applicants: Economics

Rank of Department	Applicants, %	Reviewers, %
Top 10	26	35
Other ranked	29	25
Unranked and nonacademic	44	40
TOTAL	99	100

Rank of Applicants	Rank of Department of Reviewers, %			
	Top 10	Other Ranked	Unranked and Nonacademic	Total
Top 10	36	24	40	100
Other ranked	42	23	36	101
Unranked and nonacademic	30	28	43	101

Rank of Applicants	Rank of Department of Reviewers, Mean Ratings Given			
	Top 10	Other Ranked	Unranked and Nonacademic	Total
Top 10	1.70	1.71	2.08	1.85
Other ranked	2.54	2.64	2.82	2.67
Unranked and nonacademic	2.68	2.90	3.19	2.96
TOTAL	2.35	2.54	2.80	2.59

TABLE 8 Rank of Department of Reviewers by Rank of Department of Applicants: Fluid Mechanics

Rank of Department	Applicants, %	Reviewers, %
Top 10	29	28
Other ranked	45	32
Unranked and nonacademic	26	40
TOTAL	100	100

	Rank of Department of Reviewers, %			
Rank of Applicants	Top 10	Other Ranked	Unranked and Nonacademic	Total
Top 10	27	31	42	100
Other ranked	26	38	36	100
Unranked and nonacademic	35	23	42	100

	Rank of Department of Reviewers, Mean Ratings Given			
Rank of Applicants	Top 10	Other Ranked	Unranked and Nonacademic	Total
Top 10	2.62	2.47	1.54	2.12
Other ranked	2.55	3.35	2.70	2.90
Unranked and nonacademic	3.50	3.57	2.76	3.19
TOTAL	2.86	3.14	2.37	2.75

conclude that, in algebra, top-ranked reviewers are more lenient, in general, than reviewers from other departments. This is the case in only 4 of the 10 programs: algebra, chemical dynamics, economics, and solid-state physics. In the other six programs, top-ranked reviewers were in general less lenient than were other reviewers.

The last part of the table shows whether there was any significant interaction effect. That is, are top-ranked reviewers more likely to give high scores to applicants from top-ranked departments than would be expected on the basis of the general tendency of top-ranked reviewers to give low scores and the general tendency of top-ranked applicants to get low scores? Given these two distributions, the expected mean rating of top-15 applicants in algebra would be 1.68, and the observed mean rating was 1.98. Thus, there is no evidence that old boyism is at work in this program. In fact, the data show that, if anything, top-ranked people are tougher on their colleagues at top-ranked institutions than would be expected.

The last column of Table 12 shows that only in biochemistry was there any statistically significant interaction. That is, only in biochemistry are reviewers from top-ranked departments more likely to give favorable ratings to applicants from top-ranked departments than would be expected by chance. This could indicate some degree of bias or that in this field it is possible that reviewers in top-ranked departments are more discriminating and are more able to assess high-quality proposals. In seven of the programs, the relationship had effects opposite to those expected; that is, top-ranked reviewers gave lower scores to proposals from top-ranked applicants than would be expected by chance. In the two other programs, anthropology and meteorology, the differences were not statistically significant. On the basis of these data, there is very little evidence that reviewers were biased in evaluating the proposals of their colleagues.

We also considered the effect of the geographic location of reviewers on how they evaluated proposals from applicants in different geographic locations. The results are presented in Table 13. The first column presents the mean rating given where the geographic location

TABLE 9 Rank of Department of Reviewers by Rank of Department of Applicants: Geophysics[a]

Rank of Department	Applicants, %	Reviewers, %
Top 10	36	32
Other ranked	36	31
Unranked and nonacademic	28	37
TOTAL	100	100

	Rank of Department of Reviewers, %			
Rank of Applicants	Top 10	Other Ranked	Unranked and Nonacademic	Total
Top 10	31	36	34	101
Other ranked	37	27	37	101
Unranked and nonacademic	26	31	42	99

	Rank of Department of Reviewers, Mean Ratings Given			
Rank of Applicants	Top 10	Other Ranked	Unranked and Nonacademic	Total
Top 10	2.37	2.17	2.25	2.26
Other ranked	2.23	2.69	2.05	2.29
Unranked and nonacademic	3.29	2.57	2.61	2.75
TOTAL	2.49	2.44	2.29	2.40

[a]For this table the rank of department of geophysicists is based upon the 1971 ACE ratings of geology departments.

TABLE 10 Rank of Department of Reviewers by Rank of Department of Applicants: Meteorology[a]

Rank of Department	Applicants, %	Reviewers, %
Top 17	25	15
Other ranked	33	39
Unranked and nonacademic	42	46
TOTAL	100	100

	Rank of Department of Reviewers, %			
Rank of Applicants	Top 17	Other Ranked	Unranked and Nonacademic	Total
Top 17	19	33	47	99
Other ranked	15	37	49	101
Unranked and nonacademic	14	43	43	100

	Rank of Department of Reviewers, Mean Ratings Given			
Rank of Applicants	Top 17	Other Ranked	Unranked and Nonacademic	Total
Top 17	2.79	2.98	2.94	2.45
Other ranked	2.42	2.57	2.37	2.92
Unranked and nonacademic	3.34	2.92	2.97	3.00
TOTAL	2.90	2.82	2.74	2.80

[a]Ranks based upon judgment of program director.

of the applicant and reviewer are the same. The second column shows the mean rating given when the geographic location of the applicant and reviewer are not the same. If the number in column 1 is higher than the number in column 2, there is no evidence that reviewers are, in general, more likely to favor people from the same part of the country. For 7 of the 10 fields, the relationship produces effects opposite to those expected; that is, reviewers are more harsh on proposals from people in their own areas than they are on proposals from people in other areas. In the three other areas, although the relationship is as expected, it is statistically nonsignificant.

We then tested four separate hypotheses related to whether reviewers were likely to favor applicants from their own areas. For example, the mean ratings given to applicants from the northeast by reviewers from the northeast (column 4) can be compared with the mean ratings given to applicants from the northeast by reviewers from other sections of the country (column 5). Once again, the number in column 4 would have to be lower than the number in column 5 to demonstrate regional bias in reviewing. We do find a statistically significant relationship in 2 of the 10 programs, fluid mechanics and meteorology. In these two programs, reviewers from the northeast are more lenient on proposals from applicants from the northeast than are reviewers from other sections of the country.

TABLE 11 Rank of Department of Reviewers by Rank of Department of Applicants: Solid-State Physics

Rank of Department	Applicants, %	Reviewers, %
Top 10	15	19
11-20	18	9
21-60	25	32
Unranked and nonacademic	42	40
TOTAL	100	100

	Rank of Department of Reviewers, %				
Rank of Applicants	Top 10	11-20	21-60	Unranked and Nonacademic	Total
Top 10	17	12	29	42	100
11-20	31	7	23	39	100
21-60	20	10	35	35	100
Unranked and nonacademic	14	8	35	44	101

	Rank of Department of Reviewers, Mean Ratings Given				
Rank of Applicants	Top 10	11-20	21-60	Unranked and Nonacademic	Total
Top 10	2.05	2.56	1.98	1.80	1.98
11-20	1.79	1.83	1.86	1.87	1.84
21-60	2.12	2.21	2.28	2.01	2.15
Unranked and nonacademic	2.14	2.35	2.53	2.51	2.45
TOTAL	2.02	2.28	2.30	2.17	2.20

Proposals from southerners show no evidence of any regional bias in any of the 10 programs. They are given the same evaluations by southerners as by reviewers from other sections of the country. In all, we made 50 such geographic comparisons. In only six cases did we find statistically significant differences. There is very little evidence that reviewers in certain geographic locations rate the proposals of applicants in those locations more favorably than do reviewers from other sections of the country.

For one field, biochemistry, we collected data on the citations of the reviewers.[3] We found no correlation between numbers of citations of

[3]See Appendix B for a description of how citation data were collected.

TABLE 12 Rank of Department of Reviewers by Rank of Department of Applicants: Analyses of Variance

Program	Mean Rating	Mean Rating Given to "Top Group"	Statistical Significance	Mean Rating Given by "Top Group"	Statistical Significance	Expected Mean Rating of "Top Group" by "Top Group"	Mean Rating of "Top Group" by "Top Group"	Statistical Significance
Algebra	2.06	1.81	$t = 1.96$, $p = 0.025$	1.93	$t = 1.99$, $p = 0.025$	1.68	1.98	–
Anthropology	2.50	2.55	–	2.66	–	2.71	2.68	NS
Biochemistry	2.56	1.96	$t = 4.09$, $p < 0.005$	2.62	–	2.02	1.56	$t = 2.6$, $p < 0.01$
Chemical Dynamics	2.24	1.94	$t = 2.05$, $p < 0.025$	2.01	$t = 2.38$, $p < 0.01$	1.71	1.70	–
Ecology	2.34	2.16	$t = 2.22$, $p < 0.025$	2.50	–	2.32	2.58	–
Economics	2.59	1.87	$t = 5.80$, $p < 0.005$	2.37	$t = 2.17$, $p < 0.025$	1.65	1.76	–
Fluid Mechanics	2.75	2.12	$t = 6.09$, $p < 0.005$	2.86	–	2.23	2.62	–
Geophysics	2.40	2.26	$t = 1.87$, $p < 0.05$	2.49	–	2.35	2.37	–
Meteorology	2.80	2.92	–	2.90	–	3.02	2.79	NS
Solid-State Physics	2.20	1.98	$t = 2.30$, $p < 0.025$	2.02	$t = 2.22$, $p < 0.025$	1.80	2.05	–

TABLE 13 Geographic Location of Reviewers by Geographic Location of Applicants: Analyses of Variance

Program	Mean Rating Where Geographic Location Is: Same for Applicant and Reviewer	Different for Applicant and Reviewer	Statistical Significance	Northeast Applicant, Northeast Reviewer	Northeast Applicant, Other Reviewer	Statistical Significance	Southern Applicant, Southern Reviewer	Southern Applicant, Other Reviewer	Statistical Significance	Midwest Applicant, Midwest Reviewer	Midwest Applicant, Other Reviewer	Statistical Significance	Western Applicant, Western Reviewer	Western Applicant, Other Reviewer	Statistical Significance
Algebra	2.14	1.95	–	1.91	1.97	NS	1.88	2.51	NS	1.79	2.01	NS	1.97	2.23	NS
Anthropology	2.61	2.30	–	1.85	2.05	NS	3.00	2.89	–	2.69	2.54	–	2.57	2.79	NS
Biochemistry	2.57	2.44	–	2.60	2.86	NS	2.38	2.74	NS	2.36	2.46	NS	2.42	2.43	NS
Chemical Dynamics	2.26	2.14	–	2.15	2.15	–	2.69	2.15	–	2.15	2.41	NS	2.00	2.25	NS
Ecology	2.32	2.40	NS	1.58	2.00	NS	2.60	2.83	NS	2.73	2.13	–	2.09	2.38	$t = -1$, $p = 0.0$
Economics	2.60	2.53	–	2.52	2.35	–	No cases	2.88	–	1.70	2.69	$t = -2.22$, $p < 0.02$	3.12	2.65	–
Fluid Mechanics	2.81	2.56	–	2.71	3.30	$t = -2.28$, $p < 0.02$	3.25	3.28	NS	2.33	2.39	NS	1.82	2.53	$t = -2$, $p < 0.0$
Geophysics	2.40	2.40	–	2.44	2.30	–	3.57	2.71	–	1.67	2.77	$t = -2.99$, $p = 0.04$	2.40	2.20	–
Meteorology	2.78	2.83	NS	2.49	3.08	$t = -2.76$, $p < 0.005$	3.50	2.88	–	2.70	2.73	NS	2.92	2.53	–
Solid-State Physics	2.19	2.21	NS	1.99	2.17	NS	2.60	2.41	–	2.10	2.18	NS	2.43	2.00	NS

reviewers and numbers of citations of applicants. The program director was not more likely to assign the proposals of eminent biochemists for review by other eminent biochemists, at least as eminence is measured by citations. We also found no evidence that reviewers with many citations were more likely to be lenient than were reviewers with fewer citations. As expected, however, applicants with relatively large numbers of citations to their recent work, in general, received significantly more favorable reviews than applicants with relatively few citations. Finally, and most importantly, the interaction effect is statistically insignificant. Thus, there is no evidence that highly cited reviewers are excessively favorable to the proposals of highly cited applicants.

In conclusion, we find little evidence that the characteristics of reviewers interact with the characteristics of applicants so as to influence substantively the outcome of decisions. Different types of reviewers seem to evaluate proposals of different types of applicants in much the same way. There is very little evidence for reviewer bias or for support of an old boy hypothesis. We must collect additional data to test the other forms of that hypothesis.

SECTION 3

Influence of Characteristics of Applicants on Reviewer Ratings

We have demonstrated that the ratings received by applicants are not significantly influenced by the characteristics of the scientists doing the rating—the peer reviewers. We now examine another question. To what extent are the ratings given by peer reviewers correlated with the characteristics of principal investigators, or applicants? Four essential criteria are supposed to be applied in the evaluation of applications:

1. The quality of science described in the proposal.
2. The competence of the principal investigator to conduct the research as demonstrated by past scientific performance.
3. Available facilities.
4. Geographic and institutional distribution, all other things being equal.

We can use the data on the reviews received by the 1,200 applicants we have studied to see the extent to which favorable ratings are more likely to be received by scientists in the most prestigious institutions and by scientists who have been funded by the NSF in the past.

The quantitative analysis reported in this section has two different purposes. The first is to discover the extent to which the characteristics of NSF applicants, including measures of their past scientific research performance, can be used to predict the ratings their proposals receive from peer reviewers. The second is to discover the extent to which different types of applicants with distinctly different characteristics are more or less likely to receive relatively high ratings from peer reviewers.

In order to demonstrate the difference between the two questions, consider a concrete example. The first question, relating to prediction, asks: Considering the entire sample of applicants, how well do the numbers of citations to applicants' work predict the ratings of their proposals? The second question, involving comparison of vastly different types of scientists, asks: Are the applicants whose work has received citations in the top 5 percent of the distribution of citations more apt to receive excellent or very good ratings than scientists whose work has received citations in the bottom 10 percent?

In order to answer these two questions, we have employed several analytic techniques. We have analyzed the same data using scattergrams, contingency tables, and linear regression methods. All three methods of analysis are useful in answering the first question.[1] We shall show below that each method yields the same substantive conclusions. In answering the second question we depend upon tabular analysis of contingency tables.

To discover the extent to which the characteristics of applicants are taken into account by reviewers, we have used as our unit of analysis individual pairs of applicants and reviewers. Thus, for each applicant we have two to eight different pairs. For each of the 10 programs we have approximately 300-500 cases. This enables us to see how characteristics of particular types of applicants are correlated with the ratings given their proposals by reviewers. All 10 programs we studied employed the same rating scale in which "excellent" was equivalent to one, "very good" equivalent to two, "good" equivalent to three, "fair" equivalent to four, and "poor" equivalent to five. Some programs allowed reviewers to give a score between two numbers. The distribution of ratings for each of the 10 programs is presented in Table 14.

Obviously, we do not have any measure of a crucial variable—the quality of the research proposal. We do, however, have measures of the past track records of research performance of principal investigators and data on their institution and geographic locations. It is reasonable to presume that at least part of the variance in reviewer

[1]We must point out, however, that linear regression analysis assumes that the dependent variable is a normal distribution function evaluated as a linear function of the independent variables. We do not have enough data to test whether these assumptions are valid throughout the study. In fact, some of the tables show that the linearity assumption does not hold. In these cases the tables give more information than probit. (See Tables 23 and 54 and the discussion of them in the text.)

TABLE 14 Frequency Distribution of Ratings for Each Program

Algebra

Value	Frequency	Percent[a]
1.0	78	25
1.5	20	6
2.0	122	39
2.5	19	6
3.0	58	18
3.5	2	1
4.0	10	3
4.5	1	0.3
5.0	5	2
Total	315	99
	$\bar{X} = 2.05$	$S^2 = 0.74$

Anthropology

Value	Frequency	Percent[a]
1.0	74	33
1.5	5	2
2.0	40	18
2.5	9	4
3.0	37	16
3.5	3	1
4.0	33	14
4.5	1	0.4
5.0	25	11
Total	227	100
	$\bar{X} = 2.48$	$S^2 = 1.90$

Biochemistry

Value	Frequency	Percent[a]
1.0	82	19
1.5	12	3
2.0	115	27
2.5	19	4
3.0	93	22
3.5	7	2
4.0	69	16
4.5	3	1
5.0	23	56
Total	423	100
	$\bar{X} = 2.55$	$S^2 = 1.32$

Chemical Dynamics

Value	Frequency	Percent[a]
1.0	84	21
1.5	26	7
2.0	139	36
2.5	23	6
3.0	71	18
3.5	15	4
4.0	22	6
5.0	12	3
Total	392	101
	$\bar{X} = 2.23$	$S^2 = 0.96$

Ecology

Value	Frequency	Percent[a]
1.0	106	26
1.5	22	6
2.0	117	29
2.5	16	4
3.0	63	16
3.5	4	1
4.0	41	10
4.5	3	1
5.0	30	8
Total	402	101
	$\bar{X} = 2.34$	$S^2 = 1.49$

Economics

Value	Frequency	Percent[a]
1.0	79	25
1.5	9	3
2.0	78	24
2.5	10	3
3.0	61	19
3.5	2	1
4.0	46	14
4.5	11	3
5.0	28	9
Total	324	100
	$\bar{X} = 2.58$	$S^2 = 1.69$

Fluid Mechanics

Value	Frequency	Percent[a]
1.0	64	19
1.5	3	1
2.0	91	27
2.5	3	2
3.0	70	20
3.5	3	1
4.0	63	19
5.0	38	11
Total	335	100
	$\bar{X} = 2.75$	$S^2 = 1.66$

Geophysics

Value	Frequency	Percent[a]
1.0	103	21
1.5	10	2
2.0	180	37
2.5	11	2
3.0	98	20
3.5	11	2
4.0	50	10
4.5	28	7
Total	491	101
	$\bar{X} = 2.40$	$S^2 = 1.23$

Meteorology

Value	Frequency	Percent[a]
1.0	74	13
1.5	13	2
2.0	141	25
2.5	28	5
3.0	144	26
3.5	19	3
4.0	89	16
5.0	54	10
Total	562	100
	$\bar{X} = 2.79$	$S^2 = 1.35$

Solid-State Physics

Value	Frequency	Percent[a]
1.0	86	17
1.5	27	5
2.0	213	43
2.5	40	8
3.0	92	18
3.5	5	1
4.0	29	6
5.0	6	1
Total	498	99
	$\bar{X} = 2.20$	$S^2 = 0.72$

[a]Because of rounding, percents may not equal 100.

ratings that is not explained by reference to the three factors we can measure must be related to the quality of the science proposed, or to lack of agreement among the reviewers.[2]

"TRACK RECORD" AND PEER REVIEW RATINGS

The first question we address is the extent to which scientists who have performed well in the past are more likely to get favorable reviews than are scientists who have not. It should be pointed out that some of the scientists who, in our data, appear to have poor track records are young scientists who have not had the opportunity to demonstrate their competence.

We had three indicators of scientists' past performances. Two of these are based upon citation counts, and one upon number of published papers. We have the total number of citations to the work published by scientists in the last 10 years.[3] Citations are being used as a rough indicator of the influence or quality of a scientist's published work. The second citation indicator includes all 1974 citations to the work of scientists published before 1965.[4] This is a rough measure of the reputations of scientists based upon work published more than 10 years ago.

We also examined the numbers of papers that scientists had pub-

[2]For descriptive purposes we present in Table 15 the means and standard deviations for the variables we have used for each of the 10 programs.

[3]Although the *Science Citation Index* lists only citations of the work of scientists on which they were sole authors or first authors, we looked up all references to coauthored papers published by scientists on which they were not sole or first authors and added those to our totals for those scientists. (We were unable to do this for anthropology and economics.) After collecting data on the citations to all work published in the last 10 years, we used a log transformation, because the distribution of citations is highly skewed. Most scientists have relatively few citations and a small number of scientists have very large numbers of citations. By using log transformations, we avoid any effects due to a few extreme cases.

[4]For this measure we do not have citations to coauthored papers on which scientists were not first authors. However, our data for papers published in the last 10 years, 1965 through 1974, showed that the total number of citations to first-authored and sole-authored papers and the total number of citations, including papers on which the authors were not first authors, were very highly correlated. In all eight fields for which we have data, the correlation was over 0.85. (Algebra, $r = 0.99$; biochemistry, $r = 0.86$; chemical dynamics, $r = 0.91$; ecology, $r = 0.97$; fluid mechanics, $r = 0.92$; geophysics, $r = 0.96$; meteorology, $r = 0.88$; solid-state physics, $r = 0.89$.) Therefore, the data on citations to work published prior to 1965 should adequately reflect the significance of older work. Citations to older work have also been treated with log transformations.

TABLE 15 Means and Standard Deviations for All Independent Variables

Independent	Algebra		Anthropology		Biochemistry		Chemical Dynamics		Ecology		Economics		Fluid Mechanics		Geophysics		Meteorology		Solid-State Physics	
Variables	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Citations in 1974 to work published 1965-1974 (log)	0.36	0.47	0.40	0.47	1.49	0.57	1.53	0.54	0.81	0.63	0.60	0.59	0.61	0.57	1.03	0.67	0.78	0.57	1.28	0.64
Citations in 1974 to work published before 1965 (log)	0.12	0.31	0.10	0.29	0.54	0.64	0.50	0.70	0.20	0.42	0.16	0.40	0.21	0.38	0.26	0.46	0.19	0.40	0.47	0.59
Number of papers published 1965-1974 (log)	0.80	0.45	0.97	0.69	1.30	0.42	1.35	0.41	0.87	0.50	1.05	0.76	0.83	0.56	1.08	0.52	1.0	0.46	1.28	0.42
Rank of current department	3.71	1.32	3.69	1.64	4.65	1.43	4.09	1.66	3.31	1.66	3.21	1.75	3.11	1.59	2.94	1.21	2.98	1.41	3.12	1.75
Rank of Ph.D. department	1.77	0.73	1.75	0.57	2.14	0.80	1.83	0.77	1.89	0.90	1.76	0.78	2.03	0.77	2.24	1.02	2.49	1.00	1.85	0.83
Ph.D.-granting institution/other	0.83	0.37	0.68	0.47	0.71	0.45	0.90	0.30	0.68	0.47	0.71	0.45	0.90	0.29	0.83	0.38	0.59	0.49	0.86	0.34
Professional age	1.27	0.44	1.41	0.48	1.16	0.37	1.17	0.38	1.30	0.45	1.33	0.45	1.26	0.43	1.25	0.43	1.19	0.38	1.13	0.33
Academic rank	5.08	0.86	4.89	1.09	4.78	1.43	5.12	1.05	4.84	1.31	5.11	1.15	4.75	1.38	4.74	1.52	4.91	1.32	5.14	1.03
Past NSF funding history	1.53	1.91	0.63	1.28	1.67	1.78	1.31	1.65	1.33	1.70	0.85	1.49	0.91	1.30	2.21	2.05	1.53	1.68	1.25	1.63
Mean rating	2.09	0.69	2.52	1.04	2.62	0.87	2.27	0.77	2.34	0.84	2.58	1.15	2.79	0.98	2.42	0.78	2.84	0.83	2.19	0.62
Rating of panel	–	–	2.54	0.86	2.23	0.58	–	–	2.24	0.60	3.43	1.03	–	–	–	–	–	–	–	–

lished in the last 10 years. These data were collected from the source index of the *Science Citation Index* (SCI). Since social science journals are not included in the SCI, and since the *Social Science Citation Index* began in 1972, we collected data on publications for social scientists from vitaes included in proposal jackets. For social scientists we constructed a productivity index.[5] These three measures are direct indicators of the amounts of scientific work produced by scientists in the past and the "quality" of that work as judged by the scientists' peers. There is now a large body of literature that demonstrates that citations are highly correlated with other measures of the quality of scientific work.[6] Although these three measures are not themselves direct indicators of locations of scientists in the stratification system of their disciplines, they may be used as indirect indicators. We have found (Cole and Cole, 1973) that scientific output as measured by both the number of publications and the citations to those publications is strongly correlated with the visibility of scientists, that is, the extent to which they are known and the evaluations of their work by other scientists, and the receipt of prestigious positions and of prestigious awards like the Nobel Prize and membership in the National Academy of Sciences.

We would expect that scientists who have produced the most work in the past and whose work has been the most frequently cited would be the most eminent scientists in their fields. We would also expect that those scientists should get higher ratings than scientists who have produced fewer papers and have been less frequently cited. They should get higher ratings for two reasons. First, one of the stated criteria of the National Science Foundation is the competence of principal investigators. Presumably scientists who have done the most impressive work in the past should be deemed most competent to do the research proposed in their applications and, therefore, should receive higher ratings. Second, on the average, one would expect that scientists who have done the best work in the past will write proposals for better work today.

Let us turn to data bearing on our first question, concerning the predictability of ratings, given knowledge of individual characteristics of applicants. Consider first the relationship between the number of citations to the recent work of the applicant and the ratings received.

[5]For a description of this index see Appendix B. We have also used log transformations on the number of papers published, since this variable is also highly skewed.

[6]For a complete bibliography of studies using citations see any one of the annual guides published by the Institute for Scientific Information to accompany the *Science Citation Index*.

TABLE 16 Proportion of Variance Explained (R^2) on Rating by Citations to Recent Work: 10 Programs[a]

Algebra	0.06
Anthropology	0.00
Biochemistry	0.16
Chemical Dynamics	0.14
Ecology	0.01
Economics	0.08
Fluid Mechanics	0.03
Geophysics	0.07
Meteorology	0.08
Solid-State Physics	0.08

[a]Log of citations made in 1974 to work published between 1965 and 1974.

We have used simple ordinary least-squared regression analysis here. The cell entries of Table 16 present the squared zero-order correlation coefficient, which is simply the proportion of variance on the dependent variable, ratings, explained by the independent variable, citations to recent work.[7]

The higher the numbers the more variance on the ratings can be explained by citations to recent work. In all fields except anthropology, citations to recent work explain some variance on the ratings received. The most interesting fact about these data, however, is that citations to past work explain so little variance in the ratings. Even in biochemistry and chemical dynamics, in which citations explain the most variance in ratings received, they explain less than a fifth of the variance; in most fields, they explain considerably less. This means that scientists who have demonstrated their competence by publishing frequently cited papers are more likely to receive favorable ratings but that this effect is weak. In fact, the great bulk of the variance in the ratings cannot be explained by citations to recent work.

We examined not only the 10 fields separately, but also all 10 programs combined. Since the mean and standard deviations on the relevant variables differ significantly from program to program, it is necessary first to standardize separately all the data within each field before combining data on applicants from different programs. For

[7]In view of the shortcomings mentioned earlier regarding the linear model, it does not seem worthwhile to list the numerical values of the regression coefficient. Throughout this section we present only the proportion of variance explained, as at least a qualitatively meaningful indicator of the strength of the relationship studied.

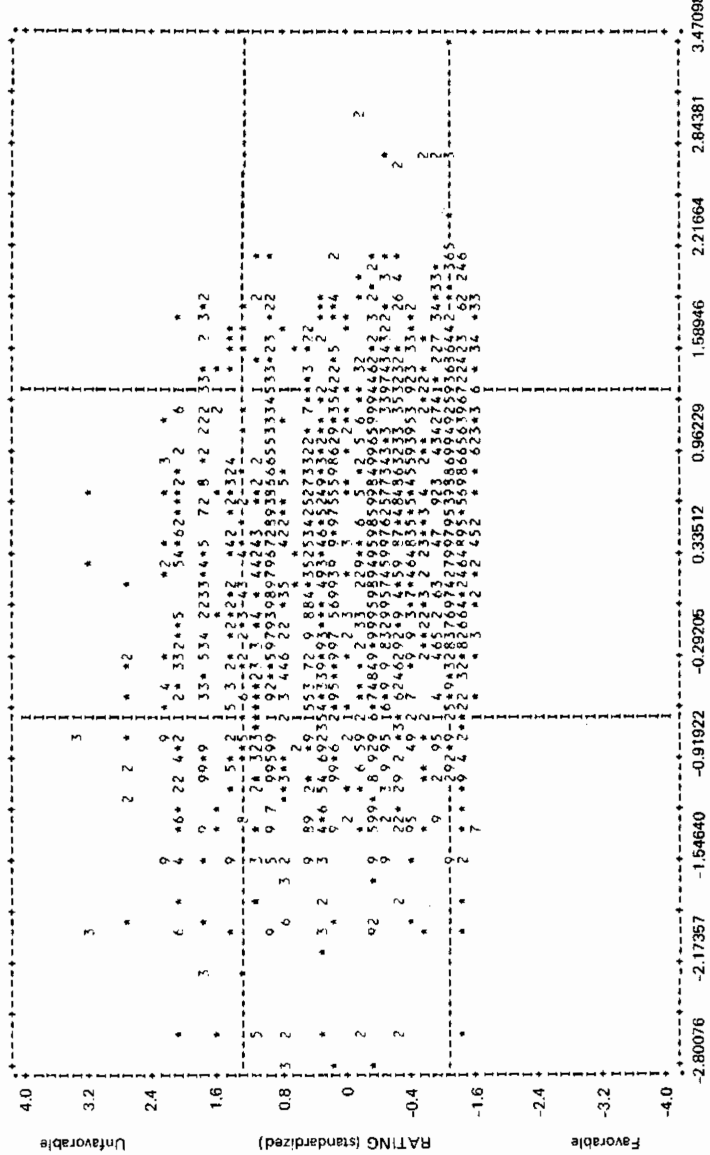

FIGURE 3 Log of citations to papers published from 1965 to 1974 and standardized peer review ratings: all 10 fields. (Note: An asterisk in the scattergram represents one case; numbers 2-8 represent that number of cases; a 9 represents nine or more cases at that point on the scattergram.)

example, we will express the number of citations received by a biochemist not in terms of an absolute number but rather in terms of the number of standard deviations above or below the mean for biochemists. Thus, a biochemist who is one standard deviation above the mean for biochemistry in citations would be treated as equivalent to an anthropologist who is one standard deviation above the mean for anthropology, despite the fact that the biochemist would have many more citations than the anthropologist. We are converting absolute scores on the variables into scores relative to other individuals in the same program. These relative or standardized scores are comparable across programs.

After standardizing the data separately within each field we were able to treat all pairs of reviewers and applicants as one sample. When we use the standardized data for all 10 fields combined, we find that citations made in 1974 to work published between 1965 and 1974 explain 6 percent of the variance in ratings.

In order to explicate still further the meaning of the results we have obtained from the regression analysis we present several scattergrams displaying the relationship between the selected variables and ratings. We begin by looking at the relationship between citations to recent work and ratings for the entire sample combined. (See Figure 3.)[8]

The cloud of points in Figure 3 indicates that there is not a strong relationship between the two variables being plotted. If citations were a good predictor of the ratings received by applicants on their proposals, we should expect that those applicants who had high citation scores, that is, were located at the far right of the scattergram, would be heavily clustered in the lower part of the scattergram, indicating that they had received "low" numerical but "high" adjectival ratings. (The reader must always invert these scores in his mind, since the NSF codes are "excellent" as a 1 and a "poor" as a 5. Between these are 2, "very good," 3, "good," and 4, "fair.") We would also expect to find those scientists who had received few or no citations, those appearing in the far left of the scattergram, clustered in the top half of the scattergram, indicating that they had received relatively low ratings on the proposals. This is clearly not the case. A substantial portion of the ratings of scientists with relatively large numbers of citations are relatively high (read low). Scientists with relatively few citations to their recent work

[8] An asterisk in Figure 3 indicates that one reviewer and applicant pair was located at this particular point in the scatterplot. Numbers 2-8 indicate the number of different pairs at those locations. The computer program used to generate these scattergrams did not have the capability of indicating a larger number than 9 at any particular location. Therefore, a number 9 indicates that 9 or more pairs were at this location in the scattergram.

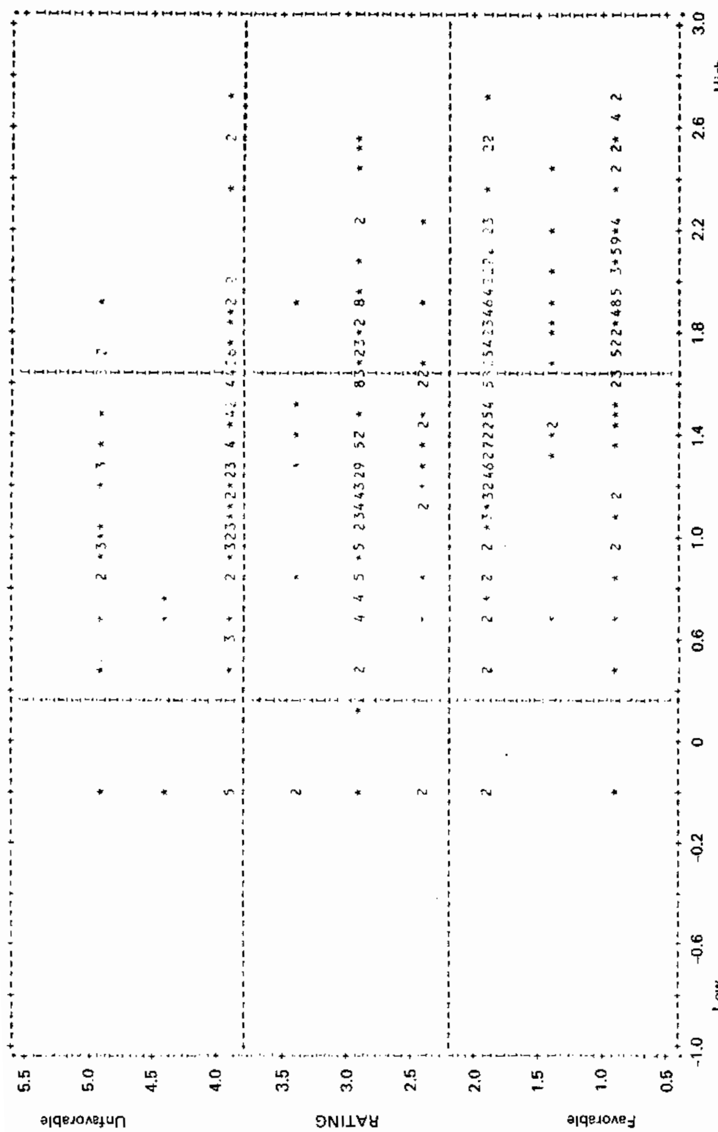

FIGURE 4 Log of citations in 1974 to papers published from 1965 to 1974 and peer review ratings: biochemistry. (Note: An asterisk in the scattergram represents one case; numbers 2-8 represent that number of cases; a 9 represents nine or more cases at that point on the scattergram.)

received relatively low (read high) ratings on their proposals. In other words, there is substantial overlap in the ratings received by highly cited scientists and those with few citations. Thus it is impossible to predict accurately the rating a scientist's proposal will get from knowledge of the number of citations to the recent work of that scientist.

We illustrate further the lack of a relationship between citations to recent work and peer review ratings by considering the results obtained for 2 of the 10 programs: biochemistry, in which the association between these 2 variables was highest among the 10 programs, and ecology, which had the second lowest association. Figure 4 presents the scattergram for the relationship between citations to recent work and ratings received in biochemistry. Again, a great number of scientists whose recent work has received a substantial number of citations obtained relatively poor peer review ratings. Correlatively, many scientists who have received few citations to their recent work obtained very good or excellent peer review ratings, represented by "low" scores on the NSF rating scale. In short, this scattergram suggests that the association between these two variables is relatively weak. This is even more apparent when we examine Figure 5, in which we present the same relationship for applicants to the ecology program. Here we see almost no relationship between these two variables.

Thus far we have used two analytic techniques to explore the possibility that a scientist's past track record is associated with peer review ratings. At least for this one indicator of track record, we have concluded that there is no substantial relationship between ratings and citations to recent work. In fact, using simple regression models we find a very pronounced lack of fit between the data and the model. Examination of the scattergrams suggests why the regression model does not provide a good description of the relationship between citations to recent work and peer review ratings. It is unlikely that any simple function could describe the data presented here.

Now we compare the results obtained from regression and scattergram analyses with those obtained from tabular analysis of the same data. In Table 17 we show the relationship between the number of citations received in 1974 to work published between 1965 and 1974 and the rating received on the proposal. For purposes of tabular analysis we have dichotomized ratings into excellent or very good (the two highest rating categories), and all others. Thus, within each program we show for each citation category the proportion of applicants who received excellent or very good ratings. For example, in 134 cases in algebra the applicant had no citations. In 63 percent of these cases the applicant received a rating of excellent or very good. In tabular

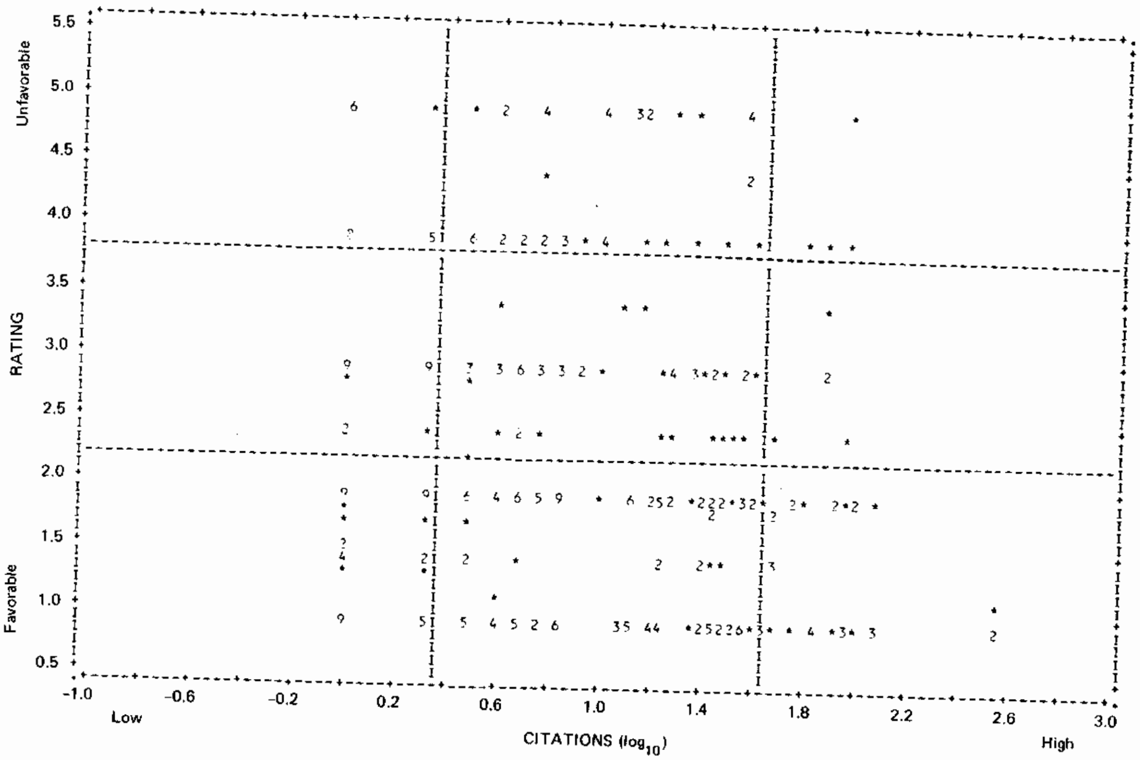

FIGURE 5 Log of citations in 1974 to papers published from 1965 to 1974 and peer review ratings: ecology. (Note: An asterisk in the scattergram represents one case; numbers 2-8 represent that number of cases; a 9 represents nine or more cases at that point on the scattergram.)

TABLE 17 Applicants Receiving Excellent or Very Good Ratings by Citations in 1974 to Work Published Between 1965 and 1974: 10 Programs

Number of Citations	Algebra, %		Number of Citations	Anthropology, %		Number of Citations	Biochemistry, %		Number of Citations	Chemical Dynamics, %		Number of Citations	Ecology, %	
0	63	(134)	0	51	(86)	0-10	23	(79)	0-11	42	(72)	0-1	65	(85)
1-2	65	(49)	1-3	46	(63)	11-22	37	(91)	12-29	49	(78)	2-4	54	(82)
3-6	72	(75)	4-8	64	(36)	23-45	45	(83)	30-56	65	(81)	5-7	59	(66)
7 or more	88	(57)	9 or more	52	(42)	46-83	61	(85)	57-81	76	(70)	8-26	54	(77)
						84 or more	80	(85)	82 or more	82	(91)	27 or more	71	(92)

Number of Citations	Economics, %		Number of Citations	Fluid Mechanics, %		Number of Citations	Geophysics, %		Number of Citations	Meteorology, %		Number of Citations	Solid-State Physics, %	
0	37	(90)	0	42	(67)	0-2	52	(99)	0-1	31	(132)	0-4	48	(103)
1	37	(35)	1-2	46	(35)	3-8	43	(98)	2-4	26	(81)	5-16	67	(112)
2-5	56	(78)	3-6	44	(81)	9-19	57	(102)	5-9	32	(111)	17-33	65	(94)
6-14	58	(59)	7-12	40	(65)	20-41	70	(91)	10-22	49	(122)	34-72	67	(100)
15 or more	68	(62)	13 or more	60	(87)	42 or more	76	(101)	23 or more	60	(116)	73 or more	83	(89)

Numbers of applicants are in parentheses.

analysis it is also necessary to categorize the independent variable, citations. We have done this by dividing up the cases within each field roughly into quintiles. Table 17 shows that in every field, scientists who have received the most citations for that field were more likely to get an excellent or very good rating on the proposal than those who were in the lowest citation category.[9] The percentage difference between high and low varies considerably from one field to another. For example, in anthropology, 51 percent of applicants in the lowest citation category received excellent or very good ratings, whereas 52 percent of those in the highest citation category received excellent or very good ratings. Biochemistry, the field in which the regression analysis indicated citations had the largest effect on ratings, shows the largest percentage difference. Twenty-three percent of those in the lowest citation category as compared with 80 percent of those in the highest citation category received excellent or very good ratings on their proposals.

The data presented in Table 17 can be used to provide answers to both of the questions we are concerned with in this section. We have already used regression and scattergram techniques in addressing the first question: To what extent can we predict the ratings of a proposal from knowledge of the number of citations to the applicant's recent work? We conclude from the tabular analysis that in the majority of fields, citations are of little or no use in predicting ratings.[10]

[9]It has been suggested that we should have drawn the dichotomy between "excellent" and all other ratings. When this was done, the results obtained in percentage differences were virtually identical with those reported in Table 17.

[10]The reader will note that we do not present tests of statistical significance for most of the tables appearing in this report. Many of the relationships shown in the tables are statistically significant. However, the reader should be aware of the difference between statistical and substantive significance. If a relationship is *statistically* significant, there is a very low probability that the percentage difference observed would be obtained by chance if there were not a real difference between groups in the population. However, statistical significance does not indicate the size or substantive significance of the difference between groups. In many of the tables that we have presented, there is a statistically significant difference, meaning that the difference displayed is unlikely to be a result of chance but nonetheless of little substantive meaning because of the small size of the difference. There is no precise way to determine whether or not a percentage difference of a particular size is substantively significant. There has been a continuing debate among sociologists for the last 20 years over the value of statistical tests of significance. (See Denton E. Morrison and Ramon E. Henkel, eds., *The Significance Test Controversy*.) This controversy is of no concern to us here. Our position here is simple: We have used these tests where we believe they add to our substantive understanding of the NSF peer review system, for example, in the analysis of variance conducted in section 2.

Solid-state physics is a field in which the relationship between citations and ratings is moderate: a total of 65 percent of the 495 ratings were either excellent or very good. If we have to guess whether or not a particular scientist would get an excellent or a very good rating, and we knew nothing else about that scientist or about his or her proposal, we would guess that he or she had a high rating and be correct in 65 percent of the cases. To what extent is this 65 percent correct prediction rate improved by knowledge of the number of citations to the applicant's recent work? The data on solid-state physics in Table 17 show that among the applicants in the lowest citation category, only 48 percent received high ratings. We would, therefore, predict that all these people would receive low ratings. We would be correct 52 percent of the time, or in 54 cases; we would be incorrect in 49 cases. In the next citation category, 67 percent of the applicants received high ratings. We would, therefore, predict that all these applicants would receive high ratings; correct in 75 cases and incorrect in 37 cases. The same arithmetic for the bottom three categories, using the number of citations received to predict rating, would give 330 correct cases and 165 incorrect cases. The proportion that we guessed correctly would be 67 percent, or only 2 percent better than what we could have done by chance without any knowledge of citations.[11]

Similar demonstrations could be done for the data on all the other fields. Biochemistry is the field in which citations have the greatest influence on ratings; they add more to our ability to predict ratings. In this field, citations enable us to predict 67 percent of the cases correctly, an increase of 17 percent over what we would have predicted by chance without any knowledge of citation. Our point here is that even though number of citations an applicant has received shows a moderate relationship with rating, citations do not add significantly to our ability to predict whether or not a particular applicant would get a high or low rating. We come to this conclusion using either tabular or regression analysis. However, we can answer a question using tabular analysis that we can't answer by regression analysis, namely: what are the conditional probabilities of receiving high ratings for scientists who differ greatly in their track records or in other characteristics we have been considering? With the tabular analysis it is possible to compare people who are at the extremes of a distribution. For example, in the programs of chemical dynamics, economics, and meteorology, applicants in the highest citation category are almost twice as likely to receive excellent or very good ratings as applicants in the lowest

[11]This is roughly equivalent to computing a lambda statistic.

citation category. However, in anthropology and ecology, there is practically no difference between the highest and lowest citation categories.

The data for the fluid mechanics program show us how the tabular data permit us to address more detailed questions than those accessible in a simple regression analysis. (See Table 17.) If we compare the fluid mechanics applicants in the lowest citation category with those in the highest citation category, we find a 20 percent difference, that is, 40 percent of applicants in fluid mechanics received no citations to their work and 60 percent of applicants in the highest citation category received either excellent or very good ratings on their proposals. These figures tell us the difference between people at the two extremes of the citation distribution. They do not tell us the extent to which the number of citations received is a good *predictor* of ratings *among all applicants* to the fluid mechanics program. In fact, if we examine the proportion receiving excellent or very good ratings on their proposals across the entire distribution of citations we find that for 80 percent of applicants in fluid mechanics, those in the first four quintiles, there is no difference whatsoever in the proportion receiving excellent or very good ratings. The only category in which there is a difference is the top category, those receiving 13 or more citations. Since citations explain no variance in rating among 80 percent of the sample, the overall predictability of ratings from citations in this field is very low.

A similar finding can be observed for solid-state physics. If we compare people in the lowest citation category with those in the highest citation category, we find a 36 percent difference. However, when we examine the 60 percent of all the applicants who fall in the three middle categories we find no difference at all in the proportion receiving excellent or very good ratings. Thus, on 60 percent of the sample in solid-state physics, citations are of no use in predicting ratings; therefore, the overall predictability of ratings from citations will be low. But, the tabular analysis shows that those scientists in the lowest citation category, at an extreme in the distribution, have a lower probability of receiving high ratings and those scientists in the top citation category, at the other extreme of the distribution, have a considerably higher probability of receiving favorable ratings on their proposals.

Table 18 presents the proportion of variance on ratings explained by citations made in 1974 to work published prior to 1965. Here we find a positive but very weak relationship in all 10 fields. Substantively, this means that scientists who are well known as a result of work published more than 10 years ago are only slightly more likely to get higher ratings than scientists who are not well known on the basis of work

TABLE 18 Proportion of Variance Explained (R^2) on Rating by Citations to Old Work: 10 Programs[a]

Algebra	0.03
Anthropology	0.02
Biochemistry	0.06
Chemical Dynamics	0.05
Ecology	0.01
Economics	0.02
Fluid Mechanics	0.07
Geophysics	0.02
Meteorology	0.01
Solid-State Physics	0.03

[a] Log of citations made in 1974 to work published prior to 1965.

published 10 or more years ago. When we use the standardized data for all 10 programs combined, we find that the citations to older work explain only 2 percent of the variance on rating. The tabular data on this variable are presented in Table 19.

Table 20 presents the proportion of variance explained on ratings by the numbers of papers published in the last 10 years. In 4 of the 10 programs (algebra, anthropology, ecology, and economics) no variance is explained. The number of papers published explains only 1 percent of variance in ratings in fluid mechanics, 7 percent for biochemistry, 9 percent for chemical dynamics, and 12 percent for meteorology. When we use the standardized data for all 10 programs combined, we find that the number of papers published in 1965-1974 explains 2 percent of the variance in ratings.

It is worth noting that there are generally high correlations between the number of papers a scientist has published and the number of times that he or she has been cited. In fact, in biochemistry, the field that on the average shows the highest correlation between the three productivity variables and the ratings, we found that all three variables together explained only 17 percent of the variance, only 1 percent more than was explained by citations to recent work.

Table 21 presents the relationship between the total number of papers published between 1965 and 1974 and ratings received by applicants in each of the 10 fields. In 7 of the 10 programs there is less than a 20 percent point difference between scientists who have published no papers and those who have published the most papers in the proportion of applicants receiving excellent or very good ratings on

TABLE 19 Applicants Receiving Excellent or Very Good Ratings by Citations in 1974 to Work Published Prior to 1965: 10 Programs

Number of Citations	Algebra, %		Number of Citations	Anthropology, %		Number of Citations	Biochemistry, %		Number of Citations	Chemical Dynamics, %		Number of Citations	Ecology, %	
0	69	(266)	0	50	(188)	0	42	(171)	0	58	(218)	0	59	(261)
1 or more	76	(49)	1 or more	59	(39)	1-6	45	(124)	1-18	64	(101)	1 or more	64	(141)
						7 or more	64	(128)	19 or more	78	(73)			

Number of Citations	Economics, %		Number of Citations	Fluid Mechanics, %		Number of Citations	Geophysics, %		Number of Citations	Meteorology, %		Number of Citations	Solid-State Physics, %	
0	49	(254)	0	43	(188)	0	55	(313)	0	39	(350)	0	61	(241)
1 or more	59	(70)	1 or more	52	(147)	1,2	65	(46)	1,2	41	(117)	1-12	67	(150)
						3 or more	70	(132)	3 or more	47	(95)	13 or more	74	(107)

Numbers of applicants are in parentheses.

TABLE 20 Proportion of Variance Explained (R^2) on Rating by Papers Published between 1965 and 1974: 10 Programs[a]

Algebra	0.00
Anthropology	0.00
Biochemistry	0.07
Chemical Dynamics	0.09
Ecology	0.00
Economics	0.00
Fluid Mechanics	0.01
Geophysics	0.04
Meteorology	0.12
Solid-State Physics	0.04

[a] A log transformation has been used.

their proposals. The exceptions are biochemistry (35 percent), fluid mechanics (26 percent), and chemical dynamics (30 percent). Note that in ecology those scientists who have published the most papers are less apt to get favorable ratings on their proposals than those who have published the least papers, and in economics there is no difference between the extreme publication categories.

It is clear that our prior expectations as to which scientists would be most likely to get high ratings on their proposals are only weakly supported by the data. If reviewers are being influenced at all by the past performances and reputations of principal investigators, the influence is not great.

These data also lead to another, more puzzling, conclusion—that there will be a low-to-moderate correlation between the perceived quality of the science in proposals submitted and the past performances of the principal investigators as indicated by published papers and citations. This conclusion is indicated because if there were a high correlation between the perceived quality of proposals and the characteristics of their authors, we would then expect to find a higher correlation between the characteristics of authors and the ratings received. There are two possibilities. One, reviewers could be basing their ratings predominantly on their perception of the quality of the science in the proposals. In this case there would be only a moderate correlation between reviewers' perceptions of the quality of the science and the past performances of the principal investigators. Two, there could be a great deal of disagreement among reviewers. (See

TABLE 21 Applicants Receiving Excellent or Very Good Ratings by Number of Papers Published Between 1965 and 1974: 10 Programs

Number of Papers	Algebra, %		Number of Papers	Anthropology, %		Number of Papers	Biochemistry, %		Number of Papers	Chemical Dynamics, %		Number of Papers	Ecology, %	
0-2	63	(68)	0-1	54	(56)	0-9	30	(80)	0-11	44	(77)	0,1,2	76	(75)
3-7	78	(85)	2-14	44	(70)	10-16	42	(87)	12-18	61	(72)	3,4,5	50	(92)
8-13	61	(77)	15-35	51	(45)	17-25	46	(83)	19-29	71	(82)	6-11	60	(75)
14 or more	75	(85)	36 or more	61	(56)	26-46	62	(87)	30-48	65	(75)	12-18	53	(66)
						47 or more	65	(86)	49 or more	74	(86)	19 or more	66	(94)

Number of Papers	Economics, %		Number of Papers	Fluid Mechanics, %		Number of Papers	Geophysics, %		Number of Papers	Meteorology, %		Number of Papers	Solid-State Physics, %	
0-1	49	(86)	0-2	41	(59)	0-4	48	(96)	0-3	33	(120)	0-9	51	(106)
2-15	53	(83)	3-9	50	(91)	5-10	50	(91)	4-9	24	(94)	10-17	72	(101)
16-51	53	(75)	10-19	34	(74)	11-20	54	(106)	10-15	44	(128)	18-25	70	(98)
52 or more	50	(80)	20 or more	57	(111)	21-31	70	(91)	16-27	46	(107)	26-44	70	(90)
						32 or more	75	(107)	28 or more	52	(113)	45 or more	65	(103)

Numbers of applicants are in parentheses.

discussion at end of this section.) The reviewing process could contain a large arbitrary element. If this is the case, we will find a low correlation between ratings given by the NSF reviewers and ratings given by independently chosen sets of reviewers. Phase 2 of this research project, which is currently under way, will investigate this possibility.

We are concerned with one other variable as an indicator of the past track records of principal investigators—the number of years out of the last 5 in which they have received NSF funds. Some applicants had received NSF funds in all or several of the years, whereas others had received NSF funds in none of those 5 years. Do applicants who currently are or recently have been NSF grant recipients have a greater likelihood of getting favorable ratings from reviewers? The data in Table 22 indicate that whether or not applicants are recent past recipients of NSF funds has very little influence on ratings of their current applications. In all 10 programs the proportion of variance explained by funding history is low. In one program it is 0, in two others it is 1 percent of the variance, and in two others it is only 2 percent of the variance. The greatest proportion of variance explained is in economics, but even here only 8 percent of the variance is explained by funding histories of applicants. When we use the standardized data for all 10 programs combined, we find that NSF funding history explains 3 percent of the variance on rating. Again, we conclude that recent NSF funding history has relatively little influence on the ratings received.

In Figure 6 we present a scattergram displaying the relationship between the funding history of applicants and ratings received on their current proposals. For this scattergram we have used the combined sample of standardized data for all 10 programs. As we would expect

TABLE 22 Proportion of Variance Explained (R^2) on Rating by Years Funded: 1970-1974

Algebra	0.05
Anthropology	0.00
Biochemistry	0.06
Chemical Dynamics	0.05
Ecology	0.02
Economics	0.08
Fluid Mechanics	0.01
Geophysics	0.01
Meteorology	0.02
Solid-State Physics	0.06

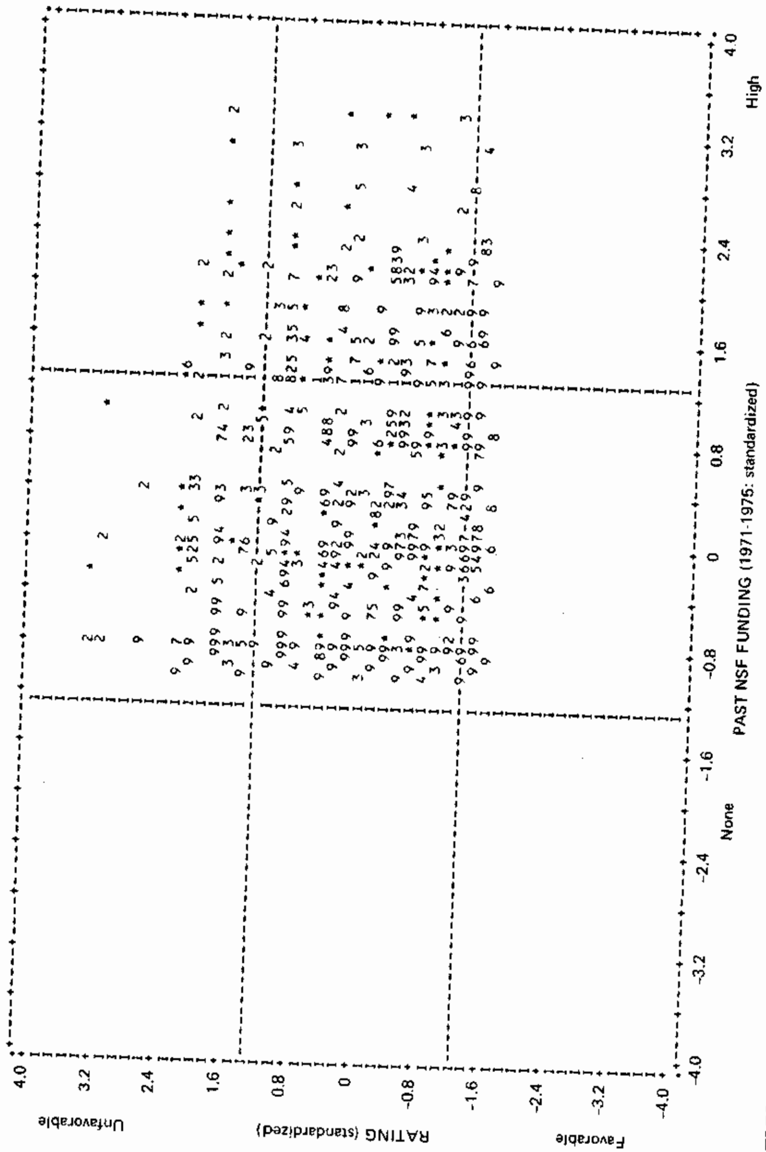

FIGURE 6 NSF funding history (1971-1975) and standardized peer review ratings: all 10 fields. (Note: An asterisk in the scattergram represents one case; numbers 2-8 represent that number of cases; a 9 represents nine or more cases at that point on the scattergram.)

from the previous associations just reported, the scattergram shows that there is no significant association between the two variables. Many applicants who have received no NSF funding in the last 5 years received favorable ratings on their proposals, and many applicants who have been funded during the entire period received unfavorable ratings on their proposals. In short, knowledge of whether or not an applicant has been funded by the NSF in the recent past is of little or no use in predicting the rating of a current proposal. It is clear from the cloud of points presented in Figure 6 why the regression results of Table 22 show no significant association between granting history and ratings received.

Table 23 shows the relationship between granting history and ratings received, using tabular analysis. For the tabular analysis we have dichotomized the applicants into those who have and those who have not received NSF funds in the last 5 years. In one program, anthropology, applicants who recently received NSF funds actually had a slightly lower probability of getting excellent or very good ratings on their proposals than did applicants who had not received NSF funds in the last 5 years. In all the other nine programs the differences between the two groups of applicants in the proportion receiving excellent or very good ratings were only slight to moderate but are definitely worth noting. The field showing the strongest relationship was economics. In this field, 73 percent of past NSF grantees and 42 percent of those who had not received NSF funds received excellent or very good ratings on their proposals.

TABLE 23 Applicants Receiving Excellent or Very Good Ratings by Past Funding History: 10 Programs

Program	Received NSF Funds in Last 5 Years, %		Did Not Receive NSF Funds in Last 5 Years, %	
Algebra	74	(152)	66	(131)
Anthropology	48	(48)	53	(169)
Biochemistry	58	(248)	37	(172)
Chemical Dynamics	74	(166)	56	(187)
Ecology	67	(181)	55	(204)
Economics	73	(95)	42	(214)
Fluid Mechanics	51	(174)	44	(151)
Geophysics	64	(297)	53	(161)
Meteorology	45	(295)	34	(232)
Solid-State Physics	76	(225)	57	(267)

Numbers of applicants are in parentheses.

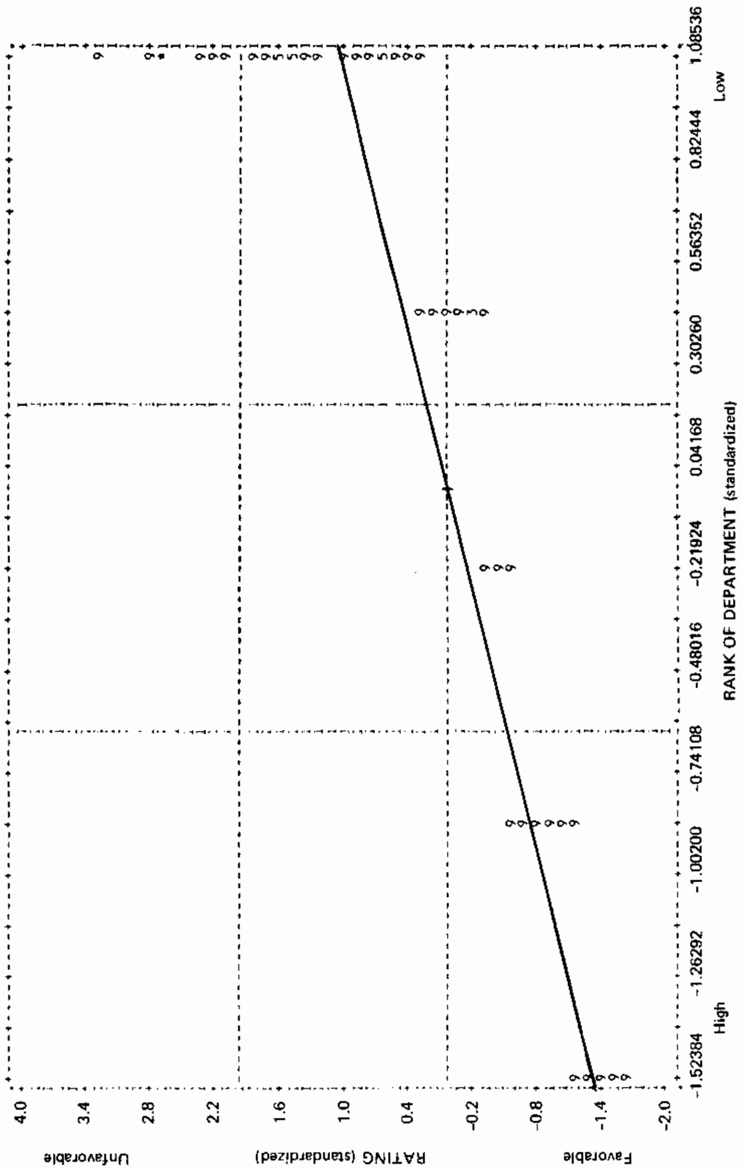

FIGURE 7 Maximum association between peer review ratings and rank of applicant's current department. (Note: An asterisk in the scattergram represents one case; numbers 2-8 represent that number of cases; a 9 represents nine or more cases at that point on the scattergram.)

LOCATION AT PRESTIGIOUS DEPARTMENTS AND PEER REVIEW RATINGS

Our data also tell us whether peer reviewers are more likely to give favorable ratings to scientists in the most prestigious academic departments. We might expect to find such a correlation, since presumably some departments are more highly ranked than others because they have more superior scientists in them. These scientists should get higher ratings both because of their capabilities as scientists and because it is presumed that their research proposals are better. As the data in Table 24 show, however, there is not a strong correlation between the rank of an applicant's current department and the rating he receives from peer reviewers. In all the programs, with the exception of anthropology, there is a correlation between the rank of applicants' departments and the ratings given their proposals; but again these correlations are surprisingly low. The greatest proportion of variance explained by department rankings is in economics, but even here only 13 percent of the variance is explained by department ranking. These data lead to the conclusion that reviewers are not being significantly influenced by the affiliations of applicants. They are only slightly more apt to give higher ratings to applicants from prestigious institutions than to those from less prestigious institutions. When we use the standardized data for all 10 programs combined, we find that rank of current department explains 5 percent of the variance in ratings.

Figure 7 shows what the relationship would be between the rank of an applicant's department and the applicant's rating if there were a

TABLE 24 Proportion of Variance Explained (R^2) on Rating by Rank of Present Department

Algebra	0.07
Anthropology	0.00
Biochemistry	0.07
Chemical Dynamics	0.02
Ecology	0.02
Economics	0.13
Fluid Mechanics	0.10
Geophysics[a]	0.03
Meteorology[a]	0.05
Solid-State Physics	0.08

[a]Rank of department scores based upon survey of NAS members (see Appendix B).

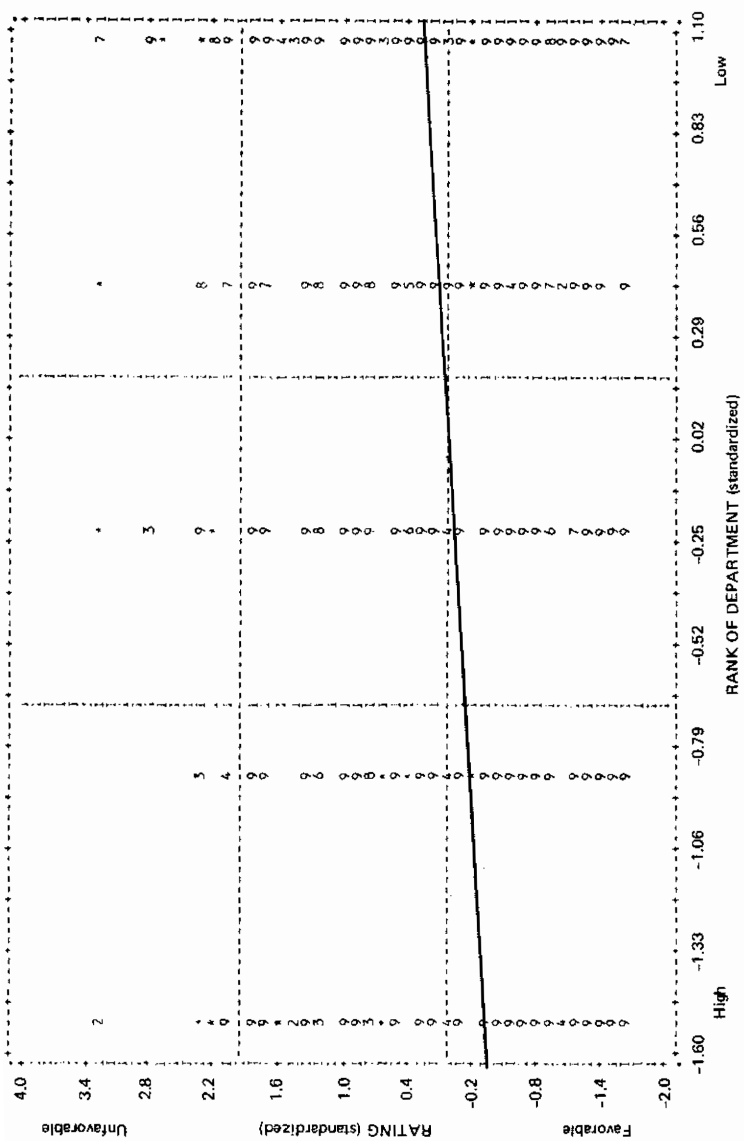

FIGURE 8 Observed relationship between peer review ratings and rank of applicant's current department. (Note: An asterisk in the scattergram represents one case; numbers 2-8 represent that number of cases; a 9 represents nine or more cases at that point on the scattergram.)

perfect or maximum association between these two variables. For this illustration we use the standardized data from the combined sample. We have roughly drawn in the least-squares regression line. The slope is 0.904. (Since the data are standardized, the correlation coefficient is equal to the slope.) Even if there were a "perfect" relationship between rating and rank of department, we would not get a correlation coefficient of 1 because rank of department is not a continuous variable. (We divided departments into five categories.) If all the highest ratings were assigned to the applicants from the highest-rank departments, only 82 percent of the variance in ratings would be explained.

The actual distribution of the data is presented in Figure 8. The slope, or the correlation coefficient, for this regression line is 0.213. Thus, the rank of an applicant's department explains only about 0.045 percent of the variance on ratings; the regression equation is an inadequate predictor of an applicant's rating. Examining the extent to which the points in the scattergram are spread out both above and below the regression line and the tremendous amount of overlap in the rating scores for people in departments of different rank emphasizes our point. Table 25 shows the relationship between the rank of an applicant's current department and the ratings received on the proposal. With the exception of anthropology, scientists who come from the highest-ranked departments are indeed more likely to receive higher ratings than those who come from unranked or nonacademic departments. However, in several of the programs, such as chemical dynamics, ecology, and geophysics, this relationship is very weak. In several of the programs the relationship is nonlinear. For example, in chemical dynamics, 78 percent of scientists from the highest-ranked departments received high ratings, 67 percent of those in the next group received high ratings, but 78 percent of those in the fourth group received high ratings—the same proportion as that received by scientists in the highest-ranked departments. Other fields show a similar lack of linearity. For example, in solid-state physics, scientists located in the lowest-ranked departments received just about the same proportion of high ratings as those in the most prestigious departments.

These findings on the relationship between rank of department and ratings seem to contradict common sense. On closer examination, however, they corroborate the findings of prior empirical studies in the sociology of science. Although it is true that, *on the average*, highly prestigious departments have more productive and talented scientists, a non-negligible proportion of talented scientists are not in the most prestigious departments. Several independent studies have found that the correlation between citations to a scientist's work and the prestige

TABLE 25 Applicants Receiving Excellent or Very Good Ratings by Rank of Current Department: 10 Programs

Program	1 Current Department Ranked High, %		2 Current Department Ranked Medium, %		3 Current Department Ranked Medium, %		4 Current Department Ranked Low, %		5 Current Department Unranked or Nonacademic, %	
Algebra	87	(23)	81	(52)	76	(71)	63	(38)	61	(131)
Anthropology	50	(24)	47	(34)	30	(10)	70	(20)	53	(139)
Biochemistry	80	(46)	68	(40)	61	(23)	33	(33)	43	(281)
Chemical Dynamics	78	(37)	67	(27)	65	(86)	78	(77)	51	(167)
Ecology	67	(104)	71	(45)	64	(22)	59	(69)	54	(162)
Economics	76	(85)	60	(57)	53	(30)	0	(13)	37	(139)
Fluid Mechanics	70	(98)	51	(61)	42	(40)	28	(50)	31	(86)
Geophysics[a]	72	(97)	64	(47)	56	(171)	55	(146)	60	(30)
Meteorology[a]	57	(122)	44	(110)	41	(95)	30	(131)	32	(104)
Solid-State Physics	81	(145)	67	(54)	71	(76)	78	(9)	52	(214)

Numbers of applicants are in parentheses.
[a]Rank of department scores based upon survey of NAS members (see Appendix B).

rank of his department is 0.30 or less. (For a review of the literature on this topic see Cole and Zuckerman, 1976.) This means that quite a few scientists who have produced high-quality work are not in highly ranked departments.

When we relate the low correlation between the quality of an individual scientist's research output and the rank of his department to the concept of self-selection we can understand better the low correlation between the rank of an applicant's department and peer review ratings. If every scientist in every department applied for a grant, there would probably be a considerably higher correlation between rank of department and rating. But we know that all scientists do not apply. Applying scientists from low-ranked departments are probably the most active researchers. Whereas six mathematicians from MIT may apply for NSF funds in a given year, perhaps only one mathematician at a lower-ranked department will apply. But this one man will possibly have a national reputation comparable to those of some of his colleagues at higher-ranked departments. The relatively wide dispersion of scientific talent and the process of self-selection may well provide the explanation of the data in Table 25.

To illustrate how tabular analysis allows us to compare people at the extremes of a distribution, we have computed an index in which applicants are given scores based upon the quintiles of their citations and the quintile ranks of their current departments. A scientist in the highest-ranked department with the highest number of citations would receive a score of 10. A scientist in the lowest-ranked department with the lowest number of citations would receive a score of 2. Table 26 shows the proportion of applicants in each index category who received excellent or very good ratings. Thirty-four percent of those in the lowest index category and 80 percent of those in the highest category received excellent or very good ratings. Since 56 percent of all the ratings were very good or excellent, we could predict 56 percent correctly by chance. Using this index composed of the two independent variables that had the strongest effect on the dependent variable

TABLE 26 Applicants Receiving Excellent or Very Good Ratings by Index of Citations and Rank of Department: All Fields Combined[a]

2	3	4	5	6	7	8	9	10
34%	43%	51%	50%	56%	61%	68%	75%	80%
(294)	(425)	(482)	(646)	(561)	(485)	(326)	(320)	(230)

The first row of numbers indicates index scores. Numbers of applicants are in parentheses.
[a]Rank of department was broken down into quintiles using the Z scores.

rating, we were able to increase the number of cases we could predict correctly to 60 percent. This suggests still further that our ability to predict ratings from these independent variables is not greatly enhanced by constructing such indices. One reason why the overall predictability is not greater is that there are relatively few cases in these extremes. For example, only 6 percent of all the cases are in the highest index category and only 8 percent of all the cases are in the lowest. A great majority of the cases are in the middle index categories between 4 and 7, where the percentage difference is only 10 points. Since this distribution is not artifactual but is representative of the distribution of the scientists who applied to the 10 programs we studied at the NSF, it cannot be discounted. Since there is not a great deal of variance in the independent variables, they are of little use in making better predictions of the dependent variable. This is one reason why citations are not a strong predictor of ratings in algebra, fluid mechanics, anthropology, and economics.

However, the data displayed in Table 26 also allow us to compare scientists who are at different ends of the index combining citations and rank of department. Let us compare the probabilities of receiving excellent or very good ratings among scientists at the two extremes. Table 26 shows a 46 percentage point difference between the two groups. This substantial difference in probability does not contradict our findings of overall low predictability because we are dealing with only 14 percent of the total sample. Table 26 shows, as one would expect and hope, that scientists with a very high index are much more likely to receive high ratings than those with a very low index. However, the number of scientists between the two extremes is so large that the index has very little predictive value.

TABLE 27 Proportion of Variance Explained (R^2) on Rating by Type of Current Institution (Ph.D. or not): 10 Programs

Algebra	0.01
Anthropology	0.01
Biochemistry	0.02
Chemical Dynamics	0.00
Ecology	0.04
Economics	0.07
Fluid Mechanics	0.01
Geophysics	0.04
Meteorology	0.02
Solid-State Physics	0.01

TABLE 28 Proportion of Variance Explained (R^2) on Rating by Rank of Ph.D. Department

Algebra	0.04
Anthropology	0.02
Biochemistry	0.02
Chemical Dynamics	0.04
Ecology	0.01[a]
Economics	0.03
Fluid Mechanics	0.00
Geophysics[b]	0.02
Meteorology[b]	0.02
Solid-State Physics	0.02

[a] Relationship is negative.
[b] Rank of department scores based upon survey of NAS members (see Appendix B).

The data in Table 27 distinguish applicants currently employed in Ph.D.-granting institutions from those employed elsewhere. This has virtually no influence on ratings of proposals by peer reviewers. Thus the criticisms that the peer review system unfairly favors applicants from prestigious Ph.D.-granting institutions are not supported by these data.

The data on the rankings of the departments in which the applicants earned their Ph.D's showed the extent to which this variable was correlated with ratings given by peer reviewers. Table 28 shows that the rankings of Ph.D. departments of applicants explain very little variance in ratings received.

PROFESSIONAL AGE AND PEER REVIEW RATINGS

Some critics of the peer review system hold that young, inexperienced applicants have less chance to receive funds than their more experienced older colleagues. We have data on the ages of applicants and on the numbers of years since applicants received their Ph.D.'s, which we call their professional age. The results of this analysis are presented in Table 29. The findings are clear. In five of the programs professional age explains no variance in the ratings given by peer reviewers. In four programs professional age explained only 1 percent of the variance in ratings. These data strongly suggest that young people have just as good a chance to receive favorable ratings of their proposals as do their older, more experienced colleagues. This conclusion is supported by

TABLE 29 Proportion of Variance Explained (R^2) on Rating by Professional Age[a]

Algebra	0.00
Anthropology	0.00
Biochemistry	0.01
Chemical Dynamics	0.01
Ecology	0.00
Economics	0.00
Fluid Mechanics	0.01
Geophysics	0.01
Meteorology	0.00
Solid-State Physics	0.03

[a]Professional age was divided into two classes—those who received Ph.D.'s in the last 5 years and those who received Ph.D.'s more than 5 years ago.

the results reported in Table 30, which shows the influence of academic rank (only for those employed in academic institutions) of applicants on ratings received. A high correlation would indicate that applicants with high academic rank have a better chance of getting favorable ratings than applicants of lower rank. Once again, the proportions of explained variance are either nonexistent or very small. Apparently, full professors do not have a significantly better chance than their lower-ranked colleagues.

Table 31 presents the relationship between professional age and ratings received for each of the 10 programs, using tabular analysis. In algebra, ecology, and meteorology, applicants who have received their Ph.D.'s within the last 5 years have slightly higher probabilities of receiving excellent or very good ratings than do applicants who received their Ph.D.'s more than 5 years ago. In most of the other programs the difference in the proportion receiving excellent or very good ratings between relatively new Ph.D.'s and older Ph.D.'s is slight. The one program that shows a moderate relationship between these two variables is solid-state physics. Sixty-nine percent of scientists who received their Ph.D.'s more than 5 years ago received excellent or very good ratings, and 49 percent of those who received their Ph.D.'s within the last 5 years received excellent or very good ratings.

TABLE 30 Proportion of Variance Explained (R^2) on Rating by Academic Rank

Algebra	0.03
Anthropology	0.00
Biochemistry	0.02
Chemical Dynamics	0.00
Ecology	0.00
Economics	0.03
Fluid Mechanics	0.01
Geophysics	0.00
Meteorology	0.00
Solid-State Physics	0.03

TABLE 31 Applicants Receiving Excellent or Very Good Ratings by Professional Age

Program	Received Ph.D. More Than 5 Years Ago, %		Received Ph.D. within the Last 5 Years, %	
Algebra	68	(213)	76	(70)
Anthropology	54	(123)	51	(85)
Biochemistry	51	(359)	43	(61)
Chemical Dynamics	66	(293)	57	(60)
Ecology	61	(264)	63	(115)
Economics	52	(181)	48	(92)
Fluid Mechanics	50	(267)	38	(39)
Geophysics	62	(348)	56	(104)
Meteorology	40	(392)	42	(98)
Solid-State Physics	69	(421)	49	(59)

Numbers of applicants are in parentheses.

COMBINING THE NINE CHARACTERISTICS

We conclude our analysis of the influence of principal investigators' characteristics on reviewer ratings with Table 32. This table presents the amount of variance explained in ratings by all nine characteristics of applicants, using multiple regression analysis. The table shows that the characteristics of principal investigators on whom we have data explain only a small portion of the variance in ratings in all 10 programs.

Economics is the program in which the largest proportion of variance in ratings—21 percent—is explained by the combination of nine characteristics of the principal investigators. We should point out that we do not know the extent to which even this variance in ratings is a result of the influence of these nine characteristics of applicants and how much is due to an unknown correlation between the characteristics of appli-

TABLE 32 Proportion of Variance Explained on Ratings Given by All Nine Characteristics of Principal Investigators (10 Variable Regression Equations[a] for Each Program)

Program	Proportion of Variance Explained in Each Program, Multiple R^2
Algebra	0.17
Anthropology	0.04
Biochemistry	0.20
Chemical Dynamics	0.16
Ecology	0.06
Economics	0.21
Fluid Mechanics	0.17
Geophysics	0.09
Meteorology	0.14
Solid-State Physics	0.17
All 10 programs combined (standardized data)	0.11

[a] If an independent variable had a "negative" correlation with rating (i.e., eminent scientists being less likely than noneminent scientists to receive favorable ratings), the variable was omitted from the multiple regression equation for that program. The omitted variables were as follows: anthropology (rank of current department and NSF funding history), ecology (number of published papers, rank of Ph.D. department, and professional age).

cants we have been studying and the quality of their proposals. It probably involves some combination of these two factors. We tentatively conclude that a significant portion of the variance in these ratings is either a result of the perception of the quality of proposals or of a random grading process. The data we are collecting in Phase 2 of this research project may shed additional light on this important issue.

From the data presented in this section we can draw two conclusions. (1) *On the average*, the nine characteristics of principal investigators that we have studied have little effect on the ratings of their proposals. (2) The scientists at the extremes of the distribution, the very highly cited and the noncited, have significantly different probabilities of receiving excellent or very good ratings. To reiterate, the tabular data show that scientists with the most citations among applicants to their programs are substantially more likely to receive favorable ratings on their proposals than those with few or no citations. The reason why this difference between the extremes does not produce greater correlations and, therefore, explain greater amounts of variance is that relatively small numbers of scientists are at the extremes. Furthermore, the relatively small size of our samples prevents us from examining in greater detail the ratings received by the most eminent applicants to particular programs. If we were to look at the top 1 or 2 percent of applicants to NSF programs, for instance, we might find that they almost invariably do get high ratings on their proposals. Further research is needed on how the proposals of the small number of extraordinarily eminent scientists fare in the National Science Foundation peer review system.

SIGNIFICANCE OF FINDINGS

The fact that the nine characteristics explain so little variance in ratings is contrary to the expectations of many people. We must therefore consider carefully the significance of our data. First, let us consider a possible error in the methodology. We have used applicant and reviewer as the unit of analysis. Applicants who had many reviewers, of course, appeared more often than those who had fewer reviewers. It is possible that there would be less agreement among reviewers on proposals that had a large number of reviewers, since program directors typically request additional reviewers when there is disagreement among the initial set. If cases on which there is disagreement are over-represented in the sample, the correlations are artificially reduced.

Indeed, it turns out that for most of the 10 programs there is a negative correlation between the amount of agreement among the reviewers of a proposal (as measured by the variance of the ratings) and the number of reviewers of the proposal. This correlation ranged between $r = 0.00$ for chemical dynamics and $r = -0.51$ for economics. To see if this correlation had any significant influence on the results reported in the tables presented in this section, we performed an experiment in the field of ecology. We chose this program because it showed a relatively high negative correlation between degree of agreement among reviewers and the number of reviewers ($r = -0.44$) and a low correlation between citations to recent work and ratings ($R^2 = 0.01$).

We divided the ecology applicants into those who had three or fewer reviewers and those who had four or more. We then ran regression equations separately in each group. If the correlations were being reduced by over-representation of applicants with large numbers of reviewers and a low level of agreement, then the proportion of explained variance should substantially increase when we divide the applicants into groups with three or fewer reviewers and four or more reviewers. The results indicated that the proportion of variance related to applicant characteristics where the proposals had three or fewer reviewers do not differ significantly from the proportion of variance in the entire sample. They did not differ at all for the research-output measures and were only slightly higher for the "granting history" variable (years funded) and rank of current department. We may conclude that the figures presented in the tables in this section are not being significantly reduced by the possible over-representation of low-agreement cases.

We are faced with the problem of understanding the unexpectedly low correlations between characteristics of the applicants and ratings. One possible explanation would be a low level of agreement among reviewers. If, for example, an applicant with a large number of citations to past work received favorable ratings from some reviewers and unfavorable ones from others, this would yield a low or 0 correlation between citations to past work of applicants and ratings received. To what extent do the several reviewers of a given proposal agree?

To begin to estimate extent of agreement, we use the standard deviation of the reviewers' ratings. In order to estimate the amount of agreement in a given field, we computed the mean standard deviation of reviewers' ratings. The data are presented in Table 33. The mean standard deviation of reviewers' ratings varies from a low of 0.31 in algebra to a high of 0.69 in ecology and meteorology. However, in using the mean standard deviation as a measure of agreement, we must also take into account the mean rating of the reviewers. Therefore, to control for the mean rating given, we have used a coefficient of variation that is the mean standard deviation divided into the mean

TABLE 33 Consensus among Mail Reviewers: 10 Programs

Program	Mean Standard Deviation of Reviewers' Ratings	Mean of Reviewers' Ratings	Coefficient of Variation
Algebra	0.31	2.1	0.15
Anthropology	0.59	2.5	0.24
Biochemistry	0.60	2.6	0.23
Chemical Dynamics	0.42	2.3	0.18
Ecology	0.69	2.3	0.30
Economics	0.34	2.6	0.13
Fluid Mechanics	0.61	2.8	0.22
Geophysics	0.61	2.4	0.25
Meteorology	0.69	2.8	0.25
Solid-State Physics	0.35	2.2	0.16

rating. There is very little systematic variation among the 10 fields. The coefficient of variation varies from a low of 0.13 for economics to a high of 0.30 for ecology. Although these numbers are not very high, they are difficult to interpret because we do not know how much variation they represent as a proportion of the total amount of variance. We are currently using analysis of variance techniques to further investigate this important problem.

Although there might be high levels of agreement among mail reviewers, we anticipated differences in levels of agreement concerning proposals of different types of applicants. To test this assumption, we examined applicants to the biochemistry program, in which citations to recent work of applicants had a relatively high correlation with ratings received. We divided the biochemists into quintiles based upon numbers of citations to their work published between 1965 and 1974. We then compared the standard deviations separately for the bottom quintile, the top quintile, and the middle three quintiles taken together. We hypothesized that there would be more agreement on ratings for the top and bottom quintiles than for the middle group. The standard deviation for the top quintile was 0.98, the bottom quintile 1.1, and the middle group 1.1. The data clearly do not support our assumption. We must conclude that at least for applicants to the biochemistry program, there is no more agreement among reviewers of highly cited scientists than there is among the reviewers of their less-cited colleagues.

In order to eliminate the possible influence of reviewer disagreement we have computed the correlation between the mean ratings of proposals and several characteristics of the applicants. By using the mean rating (a number that has meaning only to the program director and is unknown to individual reviewers) we preclude the correlations from being lowered by disagreement. The correlations between mean ratings and individual characteristics should be substantially higher.[12] The data are presented in Table 34.

The squared correlation coefficients in Table 34 are somewhat higher than those obtained when the individual rating as opposed to the mean rating was used as the dependent variable. For example, the proportion of variance explained on ratings by log of citations to recent work for the algebra program is 0.07. The squared correlation between the mean rating received by an applicant and the log of citations to recent work for the algebra program is 0.11. We conclude, at least tentatively, that the relatively low proportions of explained variance reported in this

[12]This is because means cancel out random individual variation. Means almost invariably are more highly correlated (with any given variable) than individual scores.

TABLE 34 Influence of Selected Independent Variables on Mean Ratings (R^2)

Independent Variables	Algebra	Anthropology	Biochemistry	Chemical Dynamics	Ecology	Economics	Fluid Mechanics	Geophysics	Meteorology	Solid-State Physics
Log of citations to 1965-1974 work LCIT 6 (+)	0.11	0.01	0.27	0.25	0.02	0.08	0.02	0.14	0.16	0.16
Log of citations to pre-1965 work LCIT 3 (+)	0.02	0.03	0.12	0.08	0.02	0.01	0.07	0.03	0.04	0.07
Log of papers 1965-1974 LPAP 4 (+)	0.01	0.00	0.11	0.16	0.00	0.01	0.00	0.09	0.05	0.07
Years funded 1970-1974 YRSFUND (+)	0.08	0.00	0.11	0.08	0.03	0.06	0.02	0.02	0.04	0.12
Rank of present department RANKPRES (-)	0.10	0.00	0.12	0.04	0.04	0.23	0.10	0.07	0.11	0.16
Professional age DYRPHD (-)	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.06

section are not *primarily* a result of low levels of agreement among the reviewers of each proposal. It is still a question needing further research to determine exactly how much reviewer disagreement exists and the significance of such disagreement for the peer review process. This will be fully investigated and reported on in Phase 2.

The data in this section show that, on the average, reviewers' numerical ratings of proposals are not heavily influenced by the characteristics of the applicants. Perhaps they are more likely to be influenced by the reviewers' perceptions of the quality of science proposed.