National Science Foundation

PHASE TWO OF A STUDY

Peer Review in the National Science Foundation

PHASE TWO OF A STUDY

JONATHAN R. COLE
Professor of Sociology
Columbia University

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

with the
Committee on Science and Public Policy
National Academy of Sciences

NATIONAL ACADEMY PRESS
Washington, D.C. 1981

Library of Congress Catalog Card Number 81-83123

International Standard Book Number 0-309-03182-6

Available from

NATIONAL ACADEMY PRESS
2101 Constitution Avenue, N.W.
Washington, D.C. 20418

Printed in the United States of America

April 30, 1981

Dr. John B. Slaughter
Director
National Science Foundation
1800 G Street, N.W.
Washington, D.C. 20550

Dear Dr. Slaughter:

I am pleased to transmit "Peer Review in the National Science Foundation: Phase II of a Study," conducted by Jonathan R. Cole and Stephen Cole with the Committee on Science and Public Policy of the National Academy of Sciences, in fulfillment of Contract PRM 7522600 between the Academy and the Foundation. These studies of peer review were initiated in July 1975. The report of Phase I was submitted on November 7, 1978.

Together the two parts constitute a scholarly, searching inquiry into the peer review process as it functions at the Foundation. Phase I, a statistical retrospective examination, indicated that the system functions much as one might have hoped, in the sense that no evidence was found that any specific bias but that toward excellence operates in the peer review system. The reviewers do not constitute an "old boy" network that gives advantage to its members, to investigators at any category of institutions, or to the well-established investigator generally.

Still, the specific reviewers for each given proposal are chosen by the same NSF staff who subsequently function as decision-makers, and they and the reviewers are aware of the identities of the applicants. In Phase II, COSPUP arranged to have a set of applications reviewed, independently, by reviewers of their own choice; to the extent possible, the applications were so treated as to "blind" the reviewers to the identities of the applicants. These experiments revealed that the results of the COSPUP reviews could not be said to differ meaningfully from those of the independent NSF reviewers; it proved to be impossible to blind a large fraction of the proposals from sophisticated reviewers and, if anything, blinding appeared to be counterproductive if undertaken to enhance

the likelihood of successful application by young applicants from less well-known research institutions.

In short, together, the two parts of this study are reassuring that the peer review system in the NSF generally operates in a manner that is fair and just to the applicants. Whether, indeed, this democratically equitable arrangement also maximizes the total contribution to the advancement of science that might be possible with the available resources was not tested in these studies.

The unexpected finding was the considerable degree of disparity in judgment among reviewers, not only with respect to research proposals of middle quality but also with respect to proposals that some reviewers adjudged to be excellent or unacceptable, reflecting their tastes with respect to the substance of science itself and their judgments concerning the importance and likelihood of success of the proposed research. It was this finding, in particular, that led COSPUP to formulate the set of suggestions for enhancement of the overall operation of peer review at the Foundation that is offered in Section IV of this report.

Peer review has been regarded as both the principal process underlying decision-making at the Foundation and the assurance to the taxpayer of democratic accountability concerning the use of the funds appropriated to the Foundation. The extensive examination conducted by the Coles and COSPUP indicates that continuing confidence in this system is warranted.

Sincerely yours,

Philip Handler
President

National Academy of Sciences
2101 Constitution Avenue
Washington, D.C. 20418

COMMITTEE ON SCIENCE AND PUBLIC POLICY

STEERING COMMITTEE FOR THIS REPORT

ACKNOWLEDGMENTS

Throughout the several years of this project we have had the timely help of a great many individuals. Space prohibits specific mention of them all. We take this opportunity to acknowledge a few who were extraordinarily helpful. Albert Clogston, George Pimentel, and Harold Watts must be thanked for their yeoman service as reviewer selectors. William Spindel helped us in our attempts to develop techniques of blinding research proposals. Jack Sanderson of the NSF provided us with the materials needed to carry out the study. The more than 500 scientists who gave time to participate in this experiment through their willingness to spend time reviewing proposals deserve special thanks.

Thomas DiPrete, Judith Tanur, Gary Simon, Burton Singer, and Stephen Stigler read various drafts of the manuscript and provided us with essential statistical advice as well as useful substantive comments. Stephen Appold, Walter Bourne, James Dunn, Robert L. Froman, Jr., Pnina Grinberg, Roger Pijacki, Leonard Rubin, and Susanne Wedow contributed to the collection of data and to its computer analysis. Gloria Lebowitz and Madeline Simonson acted tirelessly as administrative assistants on the project.

Our colleagues in the sociology of science at Columbia University--Bernard Barber, Robert K. Merton, and Harriet Zuckerman--gave a great number of helpful critical comments throughout the project. Robert Green of the National Academy coordinated the entire peer-review project, and we thank Robert Hume of the Academy for his editorial work.

Several members of the Committee on Science and Public Policy should be mentioned as being particularly helpful throughout the project. Albert Clogston, Lee

Cronbach, and Jack Kiefer provided us with sound statistical and substantive advice and suggestions during the entire course of the project. Thanks also go to Philip Cohen, Greta Schuessler, and Gerald S. Schatz for reading the manuscript for COSPUP and making valuable suggestions.

I. M. Singer, more than any other individual, is responsible for getting this project under way and for seeing it through to completion. His wisdom and his critical eye are gratefully acknowledged.

Stephen Cole thanks the staff of the Center for Advanced Study in the Behavioral Sciences, Palo Alto, and the John Simon Guggenheim Foundation for providing facilities and support during the project period, and Jonathan R. Cole thanks the staff at the Center for the Social Sciences, Columbia University, for the use of its facilities throughout the project.

Jonathan R. Cole
Stephen Cole

CONTENTS

INTRODUCTION AND SUMMARY

This report concludes Phase II of a study of peer review in the National Science Foundation (NSF). In Phase I,* the peer-review process at the NSF was described in greater detail than in any previous study. The report described the formal and informal structure of the peer-review system in the basic-science sections of the NSF. It also examined the extent to which peer review at the NSF is equitable.

The reasons the Committee on Science and Public Policy (COSPUP) of the National Academy of Sciences (NAS) undertook this study are given in the preface to Phase I. In a few words, peer review† was not well understood. It was time to learn how it works and whether it works fairly, and to determine what improvements or changes might further its basic aim of supporting the best scientific research.

We briefly recapitulate the major findings of Phase I:

1. There is a high correlation between reviewer ratings and grants made (Table 25, Phase I). The scores given proposals by reviewers were the most important factor in funding decisions.
2. In the aggregate, there was not a high

*Cole, S., L. Rubin, and J. Cole. 1978. Peer Review in the National Science Foundation, Phase One of A Study. Washington, D.C., National Academy of Sciences.

†By which we mean the evaluation of proposals for research in scientific disciplines by experts in those disciplines.

correlation between grants awarded and measures of the previous scientific performance of the applicant. This is surprising, since one of the stated criteria used by the NSF in evaluating proposals is the ability of scientists to conduct the research proposed.

3. Reviewers residing in major institutions were not inclined to treat proposals from scientists at major institutions more favorably. In this respect, peer review does not serve an "old boy" network.

4. Age had no strong effect on either ratings received or the probability of receiving a grant.

5. There were low or moderate correlations between reviewer ratings (and the funding decision) and the following characteristics of the applicants: prestige rank of their current academic department, their academic rank, their geographic location, and the source of their Ph.D. training.

The results of Phase I suggest that the NSF peer-review system is indeed operating fairly, when examined with sizable samples of applicants. However, this conclusion cannot be made firm unless we know that program directors are not predetermining funding decisions by their selection of reviewers.

Phase II was designed to answer several questions: Do program directors bias the peer-review process by their selection of reviewers? Is a system of "blind" or anonymous reviewing of scientific research proposals feasible and practicable? If so, would the results differ from those of conventional review procedures? This second issue of blinding was raised by a congressional inquiry into the peer-review system at the NSF. Public Law 94-471--October 11, 1976, authorizing appropriations to the National Science Foundation for fiscal year 1977, stated in Section 2 (f):

> The Director of the National Science Foundation is authorized and directed to conduct a feasibility study of operating the peer review system used in the evaluation of grant proposals within the Foundation so as to assure that the identity of the proposer is not known to the reviewers of the proposal. Any such system shall be considered to supplement and not to supplant the peer review system in operation in the Foundation on the date of the enactment of this Act.

To answer these questions we replicated the NSF's peer-review procedures with independent reviewers not selected by the NSF. In addition, proposals were blinded--that is, identifiers were removed--for some reviewers but not for others. Whether a reviewer in our experiment received a proposal in its blinded form or in its original form was a consequence of random selection. We tried to compare our experimental outcomes, and where possible proposer and reviewer characteristics, with the actual outcomes of reviews of these proposals in the National Science Foundation, and we looked for statistical signs of substantively different patterns of treatment.

Before summarizing the primary, empirically based conclusions of this study, embodied in Sections 1-3, we give the reader a quick summary of its contents. Section 1 describes the experiment designed to answer the basic questions above. It also discusses the difficulties in blinding a proposal. The blinding procedure finally adopted is described on pages 18-19. Data on the principal investigators were collected paralleling Phase I, allowing for a comparison with Phase I (pp. 14-15).

Section 2 contrasts the peer-review ratings by NSF reviewers with those of COSPUP reviewers of non-blinded proposals. The mean ratings of each proposal by the NSF reviewers and the COSPUP reviewers appear in the scattergrams of Figures 1, 2, and 3. Next, the section addresses the fundamental issue: How much would funding decisions change if they were based on the COSPUP review (Table 7)? Reversals* are caused by reviewer disagreements in their ratings of proposals. To study the sources of these variations in ratings an analysis-of-variance model is used (Tables 9, 10, and 11). The reasons for and the implications of reviewer disagreements are then discussed.

Section 3 compares COSPUP reviews of blinded proposals with NSF reviews and COSPUP reviews of the same proposals non-blinded in a format similar to Section 2. Also discussed are the difficulty in reviewing blinded proposals and the extent to which blinding was effective in concealing the identity of the principal investigator.

We now come to the major findings of this study.

---

*Reversals of funding decisions (pp. 25 ff).

1. Phase II corroborates the findings of Phase I. The correlations between applicants' track records (and research productivity and citations to past work) and reviewer ratings of their proposals are low. Moreover, there is no evidence of bias in favor of eminent established scientists by NSF program directors (Table 5).

2. On the average, proposals that received high mean ratings from NSF reviewers received high mean ratings from COSPUP reviewers of the same proposals non-blinded (p. 22ff).

3. There is no important systematic difference between the NSF and COSPUP reviewers in any of the fields examined (p. 37). We found no evidence of systematic bias in the selection of reviewers by NSF program directors.

4. The variation in reviewer ratings of the same proposal was greater than the variation in average ratings of different proposals (p. 37).

5. Somewhere between 25 and 30 percent of NSF funding decisions could be reversed were the ratings made by another equally qualified group of reviewers. In the top quintile the percentage remains over 20 percent (Table 7). (See discussion on pages 25ff.)

6. With respect to blinding, we repeat the end of Section 3:

We found: (i) As a practical matter, it was extremely difficult to conceal the authorship of proposals. (ii) The proposals of more established applicants were more difficult to blind. (iii) Average ratings of proposals tended to be lower when reviewers did not try to guess the authors. (iv) Reviewers found blinded proposals more difficult to evaluate than non-blinded proposals. (v) Reversal rates were similar but slightly higher than in the case when NSF reviews were compared with COSPUP reviews of non-blinded proposals. (vi) More reviewers would be needed to achieve a given level of stability of average ratings for blinded proposals than for non-blinded proposals. (vii) The systematic effect of blinding is small. Few proposals would rank appreciably above or below their standing in a non-blinded competition with many reviewers. (viii) Reviewer disagreement, not blinding, accounts for nearly all the reversals in a blind/non-blind comparison.

In Section 4, COSPUP's recommendations for changes in peer review at the NSF are given. These 13 recommendations are based partly on COSPUP's involvement in this

study and its findings and partly on individual members' prior experience with the peer-review process. The recommendations are accompanied by an evaluative overview not only of the peer-review system at the NSF but also of its more general principles.

I. M. Singer
Chairman, Committee on
Science and Public Policy,
1975-1980

Section 1

THE COSPUP EXPERIMENT: DESIGN AND METHODS

The COSPUP experiment was designed to answer the basic questions raised in the introduction to this report. Three of the 10 programs included in Phase I were chosen for study: Chemical Dynamics, Economics, and Solid-State Physics. The results obtained in Phase I for these three programs were representative of the results for all 10 programs.

Design

The experiment was based on 50 proposals in each of the three programs, submitted and processed by the NSF in Fiscal Year 1976--the latest 25 proposals that had been funded by NSF and the latest 25 declined--for each program. (The actual funding percentage of NSF is close to 50.) The sample did not include proposals for continuation of multiple-year funding, fellowships, and applications not processed entirely in this funding period. The distribution of mean peer-review scores for the 50 proposals in each program closely approximated the distribution obtained for the same program in Phase I.

We wanted to replicate as far as possible the NSF review procedure, with independent reviewers not selected by the NSF. After providing copies of the 50 proposals in each field, NSF had nothing more to do with the experiment, which was conducted as follows:

1. For each program we chose a chairman of a panel to select qualified reviewers of proposals.
2. Neither the chairmen of reviewer-selection panels nor any of the participants knew which proposals had been

funded and which declined, nor had they any knowledge of the NSF ratings or decisions.

3. Each panel chairman selected 10 to 18 experts in the discipline to constitute a reviewer-selection panel, which acted, in effect, as surrogate program directors.
4. The chairman asked each prospective panelist whether he was willing to participate. There were virtually no refusals.
5. The number of reviewer selectors in each program was determined by the variety of proposal topics in the sample. Panelists were distinguished researchers in the substantive areas of the proposals. The chairman determined which proposals were sent to which reviewer selectors.
6. Before the proposals were sent to reviewer selectors, we attempted to conceal the identity of their author(s). We removed title pages, lists of references, budgets, references in the text to the past work of the principal investigators, and identifying remarks or comments in the proposals. It is, of course, hardly a foolproof method for preventing identification, especially among experts in the same field. We asked reviewer-selection panelists whether they could identify the authors of proposals they received. Some could.*

*Each of the 50 proposals was sent to two reviewer selectors. For Chemical Dynamics, 97 of the 100 forms were completed and returned. On 82 of the proposal forms no attempt was made to identify the author of the proposal. Of those who guessed, 15 guessed correctly, 3 incorrectly. No reviewer selectors picked the principal investigator as one of the suitable reviewers. In Economics, of the 99 forms returned, no attempt was made to guess on 84, while 15 guessed correctly, 1 guessed incorrectly, and 14 picked the principal investigator as an eligible reviewer. For Solid-State Physics, of the 100 forms filled out and returned, there were no guesses of authorship on 79; on 18 the author was correctly identified; on 3 the author was incorrectly identified; and on 14 the P.I. was listed as an eligible reviewer. Although reviewer selectors were not inclined to guess at authorship, if they did they were likely to be correct. There may well have been others who believed they knew who the authors of the proposals were, but who decided not to participate in the exercise.

7. Each blinded proposal was sent to two reviewer-selector panelists, who were asked to list six scientists qualified to review a particular proposal. Some named as many as 10; others only three or four. The reviewer-selection panels never met. Their nominations for qualified reviewers were made, to the best of our knowledge, independently.

8. After eliminating from the lists of proposed reviewers anyone who had reviewed a particular proposal for NSF, anyone at the same institution as the principal investigator, and the principal investigator himself, we randomly assigned each reviewer a blinded or a non-blinded proposal.*

9. Each proposal, blinded or non-blinded, was initially sent to four to seven reviewers. The number of reviewers for each proposal varied with the number nominated by the selection panel. Of course, the number of reviewers of different proposals also varies in the NSF process. Where we had difficulty obtaining reviews, we asked the selectors to choose additional reviewers.†

10. Reviewers were asked to evaluate non-blinded proposals using the criteria employed by the NSF. (A copy of the letter and proposal rating sheet sent to reviewers appears in Appendix B.) The Academy letter discussing the criteria of evaluation was identical to the one used by the NSF.

11. Reviewers of blinded proposals were asked to identify authors of proposals if they could. Whether or not they attempted this, reviewers were instructed to evaluate the proposal strictly in terms of the content of the science contained in it. They were asked to ignore past track record, even if they thought they knew the principal investigator's identity. (For a copy of the letter to the reviewers evaluating blinded proposals, the proposal rating sheet, and the standard NSF letter, see Appendix B.) Evaluation forms sent to reviewers of

*We return to the problems of blinding proposals on page 9.

†For 20 proposals in Chemical Dynamics, 12 in Economics, and 16 in Solid-State Physics, we returned to the reviewer-selector panel to obtain additional names of eligible reviewers. In a small proportion of these cases the chairman of the panel supplied the additional names.

blinded proposals contained the following three questions about the blinding process:

___ Has the removal of references from this proposal made it difficult to understand? ___Yes ___No

___ Is it more or less difficult to review a proposal on which an attempt has been made to conceal the identity of the Principal Investigator:

___More difficult ___Less difficult

___Does not influence the difficulty

___ In deciding who should get NSF grants, do you think that more weight should be given to the content of the proposal or the past research performance of the Principal Investigator?

___Content of Proposal

___Past research performance of Principal Investigator

___Both about the same

The chairmen of the reviewer-selection committees were: George Pimentel of the University of California, Berkeley,* and William Spindel of the National Academy of Sciences for Chemical Dynamics; Harold Watts of Columbia University for Economics; and Albert Clogston of Bell Laboratories for Solid-State Physics. The members of the reviewer-selection panels were:

Chemical Dynamics: Jerome Berson, Gerhard L. Closs, Stanley Cristol, Leon Dorfman, James L. Dye, Harold Kart, Kenneth Kustin, Kurt Mislow, John Ross, F.S. Rowland.

Economics: Moses Abramowitz, William Baumol, Robert Eisner, Franklin Fisher, Lawrence Klein, Lionel

*George Pimentel's work on this experiment was carried out before he became Deputy Director of the NSF. At the time, he was a member of the Committee on Science and Public Policy of the NAS and a professor at Berkeley.

McKenzie, Hugh Patrick, Joseph Pechman, Albert Rees, Paul Samuelson, George Stigler, James Tobin, Arnold Zellner.

Solid-State Physics: Boris W. Batterman, Martin Blume, Walter L. Brown, Marvin L. Cohen, Dean E. Eastman, Homer Hagstrum, Bertrand Halperin, Walter Harrison, Dennis B. McWhan, Horst Meyer, Maurice Rice, Sheldon Schultz, Yuen-Ron Shen, Albert J. Sievers, Jan Tauc, Werner P. Wolf, Peter A. Wolff, Frederick L. Vook.

Some additional facts concerning the experiment should be mentioned:

(1) All potential participants in the experiment were told that this was an experiment, that the actual funding decisions had already been made, and that their reviews would not affect the careers of the scientists involved.

(2) In the course of designing the experiment, several decisions involving alternative strategies were made that could have affected its outcome. Some of these decisions closely parallel NSF decisions made in the selection of reviewers. For example, we decided to limit the number of proposals sent to any given reviewer to three, though some reviewers were named to review four or five proposals. (It has been NSF practice to use reviewers only a few times in any given year.)

(3) As reported, any person named by a reviewer-selection panel who had reviewed the same proposal for the NSF was excluded. Some problems follow from this decision. Persons nominated as reviewers by both the NSF and our panel might be viewed as most qualified, and excluding them could seem to leave only "residual" reviewers for our appraisals. But including them would have posed other problems. The correlations between the NSF ratings and the COSPUP ratings would be artificially increased by any overlap. For any NSF reviewer nominated by our panel, we did have the adjectival appraisals and the written comments from their NSF reviews, and could add these evaluations to our set of appraisals at a later stage in the analysis. We did that and the results are reported in Section 2, page 25.*

*When NSF reviewers who were selected by our reviewer-selection panel were added to COSPUP reviewers of non-blinded proposals, there was virtually no change in the average rating.

(4) We did not have panel ratings in the experiment because we did not use panels, although they are used in some NSF programs. This limits our design in Economics, in which the panel plays an important role in reaching NSF decisions (see Phase I, pp. 109ff).

Problems of Blinding

Blinding the proposals presented special difficulties. Could we succeed in blinding a proposal without changing its contents significantly? This is by no means a trivial matter, considering the nature of proposals for scientific research. How, for example, can one blind the many proposals to continue work associated with particular scientists or laboratories?

Since a blinding procedure that eliminates all references to identifying works might alter the content of the proposal, we tried several degrees of blinding, using Chemical Dynamics as a test. William Spindel, with the aid of several faculty members at Berkeley, attempted to eliminate from proposals everything that could conceivably reveal the identity of the author. This included identifying references to prior work, the budget, and the covering page. Also eliminated were substantive parts of proposals that described research done by the principal investigator that led to the newly proposed work. Further, references to collaborations between the authors of the proposals and other investigators were deleted, as well as references to unpublished works, including private communications, and references to Ph.D. theses of the principal investigator or his students.

In a large proportion of cases, the integrity of the proposal was totally compromised; its substantive content became very unclear. Moreover, since there was substantial disagreement about what was an identifying sentence, remark, or allusion, severe blinding depended heavily upon the blinder.

After extensive review of various methods of blinding, COSPUP decided that attempts to eliminate all

possibly identifying material would not only be time consuming and costly, but would so alter proposals that the new reviewers might not be reviewing the same proposals. In light of this, a blinding procedure resembling that used by many scientific journals was adopted.

We decided to: (i) remove title pages containing the names of principal investigators and their institutions, (ii) remove lists of references, budgets, and any descriptions of facilities that might identify the institutions at which the research was to be conducted, (iii) omit any direct references to past work of principal investigators, and any other remarks that would obviously lead to identification.

Data on Principal Investigators

In addition to the peer reviews we obtained in the experiment, we collected data on the principal investigators of the proposals, paralleling the information collected in Phase I (Appendix B, Phase I volume, pp. 178-181). We gathered data on the publication records of the applicants, the citations to their work, their job locations in laboratories or academic departments, the prestige ranks of their doctoral departments, their current affiliations and the prestige ratings of those, their NSF funding histories within the past five years, their professional ages, and their academic ranks. We used these data to replicate the results obtained in Phase I with an independent sample, and to examine differences between the NSF results and the results of the COSPUP experiment.

Units of Analysis

Two basic units of analysis were used in the experiment. The first was the individual proposal. Ratings in each category were averaged; so for each proposal we had three separate mean ratings: those of NSF; those for the same version in the COSPUP experiment; and those for the blinded version.

The second unit of analysis was a pair consisting of the applicant and a rater. Again there were three sets of data: ratings by NSF reviewers, by COSPUP reviewers

of non-blinded proposals, and by COSPUP reviewers of blinded proposals.*

Pattern of Responses to the Experiment

Table 1 presents data comparing our response rates with those of the NSF for the same proposals. A total of 1,662 copies (blinded and non-blinded) were sent out for review in the three disciplines.† Our response rate was 72 percent, while NSF's rate, for the same proposals, was 78 percent. We sent an initial letter requesting cooperation with the experiment, and four follow-up letters over a period of five months to reviewers who had not responded. Finally, six months after the initial mailing, we telephoned all reviewers who had not returned the proposals. (Copies of the letters appear in Appendix B.)

Table 1 breaks the figures down by program and by blinded/ non-blinded categories. NSF obtained an 85 percent response rate in Chemical Dynamics, compared with 77 percent in the COSPUP experiment; the NSF figure was 77 percent in Solid-State Physics, compared with 72 percent in our experiment; and 69 percent compared with 67 percent in Economics. The patterns of response to blinded and non-blinded proposals were also reasonably close. There was no meaningful difference between the response rates obtained for blinded and non-blinded proposals in Chemical Dynamics and those in Solid-State Physics. For Economics, there was a higher response rate for the blinded (71 percent) than for the non-blinded (63 percent) proposals. Evidently the blinding procedure did not much influence the level of responses to requests for evaluation. However, it must be emphasized that we used minimal blinding procedures so as not to compromise the substance of the proposals. Further blinding might have had a greater effect on the response rate.

*The only information used concerning COSPUP reviewers was their institutional affiliation.

†Since some reviewers received more than one proposal, the total number of proposals sent out for review does not equal the total number of reviewers.

TABLE 1 Patterns of Response to the Peer Review Experiment for the Three NSF Programs

Program	Non-blinded Reviews	Blinded Reviews	Non-blinded & Blinded Reviews	NSF Regular[a] Review
	Type of Review Procedure			
Chemical Dynamics				
Total Sent	274	276	550	285
Total Reviewed (%)	213 (78%)	212 (78%)	425 (77%)	242 (85%)
Total Refused (%)[b]	30 (11%)	32 (12%)	62 (11%)	43 (15%)
Total Unanswered (%)[c]	31 (11%)	32 (12%)	63 (12%)	
	(100%)	(100%)	(100%)	(100%)
Economics				
Total Sent	286	281	567	235
Total Review (%)	182 (63%)	199 (71%)	381 (67%)	163 (69%)
Total Refused (%)	50 (18%)	37 (13%)	87 (15%)	72 (31%)
Total Unanswered (%)	54 (19%)	45 (16%)	99 (18%)	
	(100%)	(100%)	(100%)	(100%)
Solid-State Physics				
Total Sent	266	279	545	248
Total Reviewed (%)	190 (71%)	203 (73%)	393 (72%)	192 (77%)
Total Refused (%)	43 (16%)	39 (14%)	82 (15%)	56 (23%)
Total Unanswered (%)	33 (13%)	37 (13%)	70 (13%)	
	(100%)	(100%)	(100%)	(100%)
Combined Fields				
Total Sent	826	836	1662	786
Total Reviewed (%)	585 (71%)	614 (73%)	1199 (72%)	598 (78%)
Total Refused (%)	123 (15%)	108 (13%)	231 (14%)	170 (22%)
Total Unanswered (%)	118 (14%)	114 (14%)	232 (14%)	
	(100%)	(100%)	(100%)	(100%)

[a]Data for the NSF regular procedure may be incomplete. We have combined information on refusals and unanswered requests made by the NSF.
[b]Total Refusals refers to the reviewers who sent back proposals but refused to review them.
[c]Total Unanswered refers to the reviewers who never responded to our requests.

In summary, the rate of response to the experiment was sufficiently high to allow meaningful and valid comparisons with the NSF results. The 72 percent overall response rate is a generally acceptable level for mail surveys, and is comparable to that obtained by the NSF.

Additional data on response rates are presented in Table 2, which shows the number of reviews obtained for each of the proposals. We have subdivided the table into four columns. They give the frequency distribution in the following cases: blinded, non-blinded, blinded and non-blinded (pooled), and NSF. In Chemical Dynamics, the NSF received somewhat more reviews per proposal than COSPUP did, as would be expected from the overall response rate; in Economics, the opposite was true, perhaps reflecting that program's reliance on panels for review of proposals. In Solid-State Physics and Chemical Dynamics the COSPUP experiment had more dispersion in the number of reviewers per proposal than did the NSF. When we combine the blinded and non-blinded responses, the COSPUP experiment had, on average, considerably more reviews per proposal than the NSF.

Comparing COSPUP and NSF Reviewers

To what extent were the COSPUP reviewers similar in professional stature to the NSF reviewers? The only information we collected about our reviewers was their current institutional affiliations. This enabled us to compare the departmental rank of COSPUP reviewers with that of NSF reviewers. (See Table 3.) There was substantial overlap in this characteristic of COSPUP and the NSF reviewers, with some exceptions. In Economics, COSPUP was slightly more likely to select reviewers currently employed in the top-ranked departments (those in the highest-rated group according to the 1970 Roose-Andersen survey). COSPUP was also slightly less likely to select reviewers from either unranked departments or non-academic settings. In Solid-State Physics, COSPUP reviewers were substantially more likely to have been selected from non-academic institutions. In Chemical Dynamics, the distributions for COSPUP and NSF reviewers were very similar.

TABLE 2 Distribution of Proposals by Reviewers

Program:	Number of Reviewers	Blinded N	Blinded (%)	Non-blinded N	Non-blinded (%)	Blind & Non-blind N	Blind & Non-blind (%)	NSF N	NSF (%)
Chemical Dynamics	1							0	0%
	2	2	4%	3	6%			0	0%
	3	16	32%	4	8%			3	6%
	4	11	22%	24	48%			15	30%
	5	10	20%	15	30%	1	2%	23	46%
	6	11	22%	4	8%	6	12%	6	12%
	7					10	20%	2	4%
	8					8	16%	1	2%
	9					10	20%		
	10					6	12%		
	11					8	16%		
	12					1	2%		
Economics	1	1	2%	1	2%			8	16%
	2	4	8%	8	16%			7	14%
	3	11	22%	3	6%			16	32%
	4	16	32%	17	34%	1	2%	7	14%
	5	15	30%	8	16%	4	8%	8	16%
	6	3	6%	3	6%	10	20%	3	6%
	7					9	18%	1	2%
	8					13	26%		
	9					3	6%		
	10					7	14%		
	11					3	6%		
	12								

Program:	Number of Reviewers	Blinded N	Blinded (%)	Non-blinded N	Non-blinded (%)	Blind & Non-blind N	Blind & Non-blind (%)	NSF N	NSF (%)
Solid-State Physics	1			1	2%			0	0%
	2	1	2%	7	14%			4	8%
	3	14	28%	10	20%			11	22%
	4	21	42%	19	36%	1	2%	26	52%
	5	9	18%	9	18%	2	4%	8	16%
	6	5	10%	4	8%	11	22%		
	7					7	14%	1	2%
	8					11	22%		
	9					10	20%		
	10					3	6%		
	11					4	8%		
	12					1	2%		
Combined Fields	1	1	1%	2	1%			8	5%
	2	7	5%	18	12%			11	7%
	3	41	27%	27	18%			30	20%
	4	48	32%	60	40%	2	1%	48	32%
	5	34	23%	32	21%	7	5%	39	26%
	6	19	13%	11	7%	27	18%	9	6%
	7					26	17%	4	3%
	8					32	21%	1	1%
	9					23	15%		
	10					16	11%		
	11					15	10%		
	12					2	1%		
			(101%)		(99%)		(99%)		(100%)

TABLE 3 Rank of Academic Department for COSPUP and NSF Reviewers

Rank of Dept.	Distribution of Reviewers: Chemical Dynamics Blinded	Chemical Dynamics Non-Blinded	NSF
1	19%	25%	20%
2	19	13	14
3	14	17	14
4	11	18	14
5	10	6	12
0 unranked	18	14	17
not academic	9	7	9
	100%	100%	100%
	(212)	(213)	(242)
Economics			
1	44%	51%	37%
2	24	17	17
3	6	7	4
4	4	4	5
0 unranked	15	13	20
not academic	7	8	18
	100%	100%	101%[a]
	(199)	(182)	(163)
Solid-State Physics			
1	27%	30%	20%
2	8	11	15
3	8	7	13
4	2	3	3
0 unranked	10	10	22
not academic	46	40	27
	101%[a]	101%[a]	100%
	(203)	(190)	(192)

[a]Result of rounding.

The Pool of Eligible Reviewers

When the COSPUP experiment was originally proposed, some scientists felt that it would be difficult to carry out because the number of possible reviewers for each proposal was very limited due to intense specialization. It was suggested that the NSF would already have tapped most of the eligible reviewers. What constitutes an "eligible" reviewer, although never clearly defined by the NSF, is an important question. In our interviews with program directors, most claimed that they try to select reviewers from the general subject area of the proposal as well as from different but related areas. The latter presumably will bring broader perspective to their evaluations. In actual practice the same program director may seek broader perspectives for some proposals than for others.

The COSPUP experiment provided some data allowing us to estimate the size of the pool of reviewers for the proposals we studied. If the number of reviewers was small, a fairly high proportion of NSF reviewers would be selected by our reviewer selectors. (See Table 4.) In all three programs about 80 percent of the NSF reviewers were not selected by either of the two COSPUP reviewer selectors, about 15 percent were selected by one, and only a handful by both. Furthermore, about 10 percent of all scientists chosen as reviewers by the COSPUP selectors were chosen by both selectors.

These data indicate that the pools of eligible reviewers for most proposals number at least 10 and, given the relatively low overlap rates we found, we would predict that, if other equally qualified selectors were employed, we would find the size of the pool of eligibles to be larger than 20.

Comparison of Phase I and Phase II

It is appropriate to ask whether the sample of proposals in Phase I and that used in Phase II would produce the same general results. That is, would data based upon a sample of 1977 proposals yield results comparable to those obtained with a sample of 1975 proposals? To answer this question, we replicated a part of the analysis in Phase I on the data used in the Phase II

TABLE 4 Overlap in Choice of Reviewers by NSF and COSPUP

NSF Reviewers	Program		
	Chemical Dynamics	Economics	Solid-State Physics
Not selected by either COSPUP selector	78%	79%	83%
Selected by one COSPUP selector	18	14	17
Selected by two COSPUP selectors	4	8	1
	100%	101%[a]	101%[a]
	N (242)	N (163)	N (192)
% of COSPUP reviewers selected by two selectors	8%	10%	9%
	N (425)	N (381)	N (393)

[a]Result of rounding

COSPUP experiment. The results of one feature of this replication are presented in Table 5.*

Table 5 presents zero-order correlation coefficients between characteristics of applicants to the NSF and the individual reviewer ratings. It divides each program by the source of the reviews: NSF Phase I; NSF Phase II; COSPUP non-blinded reviews; COSPUP blinded reviews. For example, take the relationship between ratings and past funding history for applicants to the Chemical Dynamics program. In Phase I we obtained a zero-order correlation of .22; it was .21 for the NSF Phase II; and .24 for the non-blinded proposals using the COSPUP reviewers. An examination of the comparative correlations for the relationship between log of new citations and reviewer ratings shows very similar correlations. Some minor differences may be observed, but over all there is substantial similarity in the results.

Two conclusions may be drawn from the data in Table 5.

*Phase II means and standard deviations for the independent variables used in Table 5 are presented in Table 6.

TABLE 5 Correlation Coefficients of Applicant Characteristics and Reviewer Ratings: A Comparison of Phase One, NSF and Experimental Results for Three Fields[a]

Independent Variables	Chemical Dynamics				Economics				Solid-State Physics			
	NSF		COSPUP		NSF		COSPUP		NSF		COSPUP	
	Phase One	Phase Two	Non-Blind	Blind	Phase One	Phase Two	Non-Blind	Blind	Phase One	Phase Two	Non-Blind	Blind
Professional Age	.10	.09	.05	-.05	.00	-.02	-.06	-.12	-.10	-.16	-.11	.02
Rank of PhD Department	.20	.23	.24	.24	.17	-.01	-.07	-.07	.14	.23	.05	.12
Rank of Present Department	.14	.17	.26	.27	.36	.34	.28	.36	.28	.16	.24	.21
Academic Rank	.00	.15	.18	.05	.17	.16	.19	.00	.17	.00	.00	.18
Log of Papers	.30	.28	.20	.11	.00	.31	.24	.14	.20	.15	.13	.08
Log of New Citations	.37	.33	.28	.22	.28	.40	.26	.16	.28	.20	.22	.17
Log of Old Citations	.22	.15	.24	.09	.14	.21	.09	-.09	.17	.16	.15	.07
Funding History	.22	.21	.24	.20	.28	.26	.23	.14	.24	.24	.28	.29
N	(392)	(242)	(213)	(212)	(324)	(163)	(182)	(199)	(498)	(192)	(190)	(203)

[a]These correlation coefficients are based upon an analysis of data sets in which the unit of analysis was a pair consisting of a principal investigator and a single reviewer. Thus, each principal investigator would appear in the data set as many times as his or her proposal was reviewed. This analysis was aimed at understanding the influences on ratings given by particular reviewers.

TABLE 6 Means and Standard Deviations of Applicant Characteristics (Phase Two Data)

Variable	Chemical Dynamics		Economics		Solid-State Physics	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Professional Age[a]	.82	.39	.66	.48	.78	.42
Rank of Ph.D. Department[b]	1.96	1.09	1.96	1.34	1.72	.96
Rank of Present Department[c]	3.72	1.92	2.71	1.99	4.18	2.05
Academic Rank[d]	2.30	.86	2.35	.73	2.32	.79
Log of Papers[e]	1.35	.42	.90	.34	1.42	.32
Log of New Citations[f]	1.53	.54	.96	.56	1.36	.52
Log of Old Citations[g]	.70	.77	.35	.61	.53	.60
Funding History[h]	1.32	1.66	1.10	1.59	1.06	1.54
N	50		50		50	

[a]Professional age is a function of the length of time since the principal investigator's Ph.D. degree. The variable was treated as a dichotomy, with those receiving their Ph.D. degrees in 1970 or before being coded 1, and those receiving Ph.D. degrees in 1971 or later being coded as 0.

[b]The rank of the principal investigator's Ph.D.-granting department was determined from the 1964 American Council on Education survey of graduate departments. Principal investigators who received their Ph.D.'s from foreign institutions were treated as missing data. Ranks of departments were converted so that a high score on this variable represents a high ranking. American departments not ranked in the 1964 study were placed at the bottom of the scale, at zero.

[c]Rank of current department was based upon data from the 1969 American Council on Education survey of graduate education. Principal investigators not employed in academic institutions were treated as having missing data; ranks of departments were converted so that a high score on this variable represents a high ranking. American departments not ranked in the 1964 study were placed at the bottom of the scale, at zero.

[d]Full professors were coded as 3, associate professors as 2, and researchers, post-docs, assistant professors, lecturers, and instructors were all coded as 1. Principal investigators with no academic position were treated as having missing data for this variable.

[e]The total number of papers published by the principal investigator between 1967 and 1976 was determined by making a count from the source index of the relevant science citation indices. The program we used in analyzing these data had a default option, which automatically coded all zeroes as equal to 1. This number was then converted to its log to base 10.

[f]All citations to work published between 1967 and 1976 were counted in the 1976 edition of the Science Citation Index. We also counted citations to co-authored papers published in these years on which the principal investigator was not the first author. The program we used in analyzing all these data had a default option, which automatically coded all zeroes as equal to 1. The total number of citations received in this period was then converted to its log to base 10.

[g]All citations to work published before 1967 were counted in the 1976 edition of the Science Citation Index. The program we used in analyzing these data had a default option, which automatically coded all zeroes as equal to 1. This number was then converted to its log to base 10.

[h]Past funding history was coded simply as the number of years since 1972 in which the principal investigator had received NSF funds. Thus, the range on this variable was 0 to 5, with a 5 indicating that the principal investigator had received NSF funds in all the five previous years, and a 0 indicating that the principal investigator had not received any funds in the five previous years.

First, they provide corroborative evidence for the results presented in the Phase I monograph. Low correlations between an applicant's track record, research productivity, and citations to past work, and the ratings of his proposal are obtained once again with these independent ratings on a different set of proposals. Second, these results provide no evidence for bias in favor of eminent established scientists by NSF program directors.

Section 2

COMPARISON OF PEER-REVIEW RATINGS: NSF AND COSPUP REVIEWERS OF NON-BLINDED PROPOSALS

In Section 1 we described the COSPUP peer-review experiment, in which we selected independent groups of reviewers to evaluate each proposal. Our purpose, in part, was to see whether NSF program directors selected reviewers biased either for or against particular proposals. In this chapter we present an analysis of the data we obtained regarding this issue. COSPUP reviewers were asked to evaluate non-blinded proposals guided by the same set of criteria used by NSF reviewers for the identical proposals. The letters sent the experimental reviewers were exactly the same as those sent the NSF reviewers.* To what extent, then, did the two groups of reviewers reach similar conclusions?

First we compare the mean score given a proposal by the NSF reviewers and that given by the COSPUP reviewers. Second, we determine how many of the proposals would have been funded by <u>both</u> sets of reviewers or declined by <u>both</u> sets of reviewers, and in how many cases the two ratings would presumably have led to contrary funding decisions ("reversals") if the ratings were the sole basis of decision. Third, we analyze sources of differences among reviewers. Fourth, and finally, we correlate ratings with characteristics of the principal investigator. We examine, for example, whether proposals by eminent principal investigators tended to be rated higher by NSF than by COSPUP reviewers.

*Of course, COSPUP reviewers were informed that they were participating in an experiment.

Comparison of Mean Ratings

Consider first the relationship between the ratings of proposals using the NSF procedures and those using the COSPUP procedures. Recall that the reviewers in both cases were asked to evaluate proposals and give them adjectival ratings from excellent to poor. Since reviewers frequently marked the scale somewhere between two of the adjectival designations, we have magnified the scale and assigned numerical values to the adjectival designations so that the scale ranges from 50 (excellent) to 10 (poor). All appraisals that fell between scale points were assigned midpoint values; for example, if a reviewer rated a proposal between excellent and very good, we recorded a numerical value of 45.

We look initially at simple descriptive statistics for Chemical Dynamics that compare reviewer ratings made by NSF reviewers and those made by COSPUP reviewers of non-blinded proposals. The mean proposal rating is the simple average of all review scores received for a particular proposal. The number of NSF reviews entering an average may differ from the number of COSPUP reviews of a proposal.

In Chemical Dynamics, the grand mean of NSF ratings was approximately 38 on the 10-to-50 scale (i.e., between "very good" and "good"). The same proposals in the non-blinded format by COSPUP reviewers had a grand mean of 35. There was, then, little difference in the overall average ratings by the sets of reviewers; the COSPUP reviewers were slightly "harder" on the proposals. Perhaps the experimental COSPUP reviewers were a bit less favorable (or more severe) because they knew that their ratings would not affect the actual funding decisions. (However, we did not test this notion.) The standard deviations of mean ratings were also similar: 5.85 for the NSF reviews and 6.45 for the COSPUP non-blinded reviews.

The pattern is similar in the two other programs. In Solid-State Physics, the grand means were 38 for NSF and 35 for COSPUP reviewers of non-blinded proposals; in Economics, the figures were 34 and 31, respectively. The standard deviations of the two groups of reviewers were also similar: 6.15 and 6.33 in Solid-State Physics and 9.76 and 9.28 in Economics.*

*A complete list of reviewer ratings for each of the 150 proposals using the three different reviewing methods appears in Tables A-1 through A-9 of Appendix A. With these raw data the interested reader can reconstruct much of the analysis presented in this report, and can subject the data to different analytic procedures. We have, of course, concealed the identity of the principal investigators of these proposals. Identification numbers are used instead of names.

Reviewers tended to cluster their appraisals toward the top of the scale; they generally did not use its full lower range. This pattern, noted also in Phase I, implies that, in its funding decisions, the NSF must distinguish among proposals with similar average scores.

We now compare the mean NSF ratings with those from the COSPUP non-blinded reviews for Chemical Dynamics (Figure 1). A moderate degree of association appears, the correlation being $r = .60$. On the average, then, proposals that received high mean ratings from NSF reviewers also received high mean ratings from COSPUP reviewers. For most of the proposals the two mean ratings were similar: for 45 of the 50 proposals, the NSF mean score and the COSPUP mean score were within seven points of each other, on a scale whose adjectival ratings were 10 points apart.

In each scattergram, a point has been circled if either coordinate is based on only one or two reviewers. Single ratings and averages based on only two ratings tend to spread more than other averages and are less likely to be confirmed by an independent set of raters.

Figures 2 and 3 show similar results for Economics and Solid-State Physics. The correlations are .66 for Economics and .62 for Solid-State Physics. There is no important difference among the correlations. Note the 19 encircled dots in Economics. Fifteen proposals reviewed by the NSF had two or less reviewers. This reflects the fact that Economics also relies on panels to rate proposals. (See (4) on p. 9.)

The scattergrams also indicate that the two sets of ratings differ to some extent. There might be different funding decisions if these decisions were based on the COSPUP reviews. We discuss this matter next.

Reversals

We now come to a fundamental issue: How much would funding decisions change if they were wholly determined

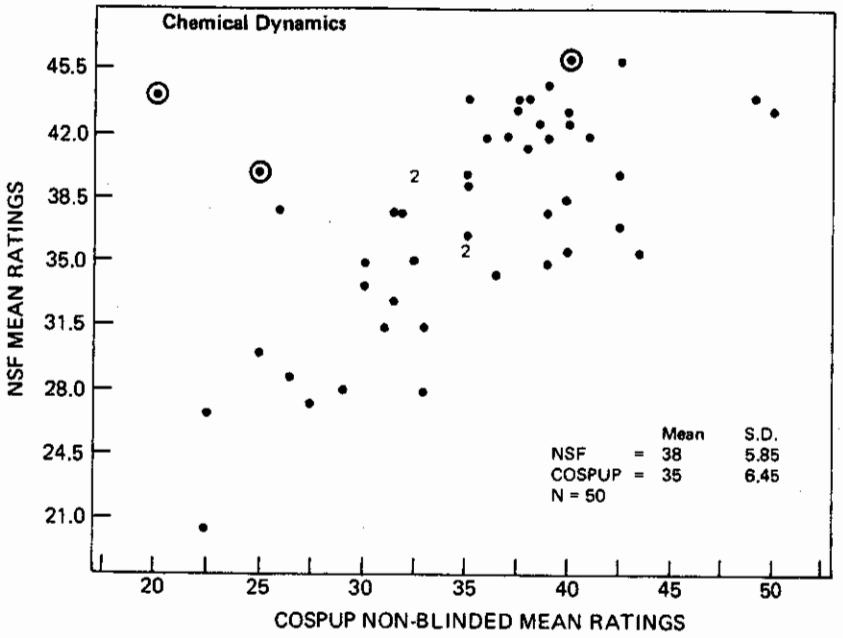

FIGURE 1 Scattergram of mean NSF ratings with mean COSPUP ratings of non-blinded individual proposals. Encircled dots indicate proposals that received two or less reviews from either COSPUP or NSF reviewers.

by the COSPUP ranking? Since many proposals with nearly identical rankings fell on opposite sides of the fund/no-fund dichotomy,* reversals are almost inevitable.

Note that we will be using NSF ratings to define reversals. That is, if a proposal was rated among the top 25 cases by COSPUP and the bottom 25 by NSF, or if rated in the bottom 25 cases by COSPUP and the top 25 by NSF, it was classed as a reversal. Since the NSF funding decision did not follow NSF ratings exactly, the findings reported below should be treated cautiously and as indicative of levels of reversals that might be obtained

*The Foundation can decide to fund a proposal at a reduced level. We still call this decision "funding" in the dichotomy.

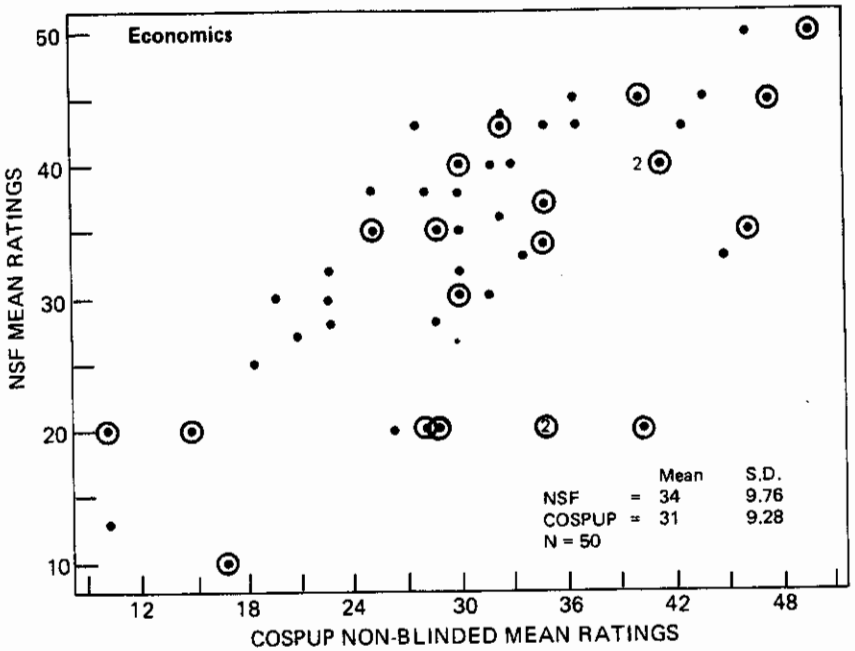

FIGURE 2 Scattergram of mean NSF ratings with mean COSPUP ratings of non-blinded individual proposals. Encircled dots indicate proposals that received two or less reviews from either COSPUP or NSF reviewers.

were a set of independently selected reviewers used. We also examine reversals when NSF actual decisions are compared with COSPUP mean ratings. So our figures on reversals are of two kinds, one for NSF mean rating and one for NSF decisions.*

We wanted to examine the extent to which reversals were concentrated at the midpoint. The COSPUP rankings for each field were grouped into quintiles, the first

*There was little difference in the proportion of reversals. The reader can judge for himself. There was no difference in Chemical Dynamics. Note that in the top quintile for Economics and Solid-State Physics the percentage of reversals would decrease noticeably if COSPUP (NB) were compared with NSF decisions. See Table 7.

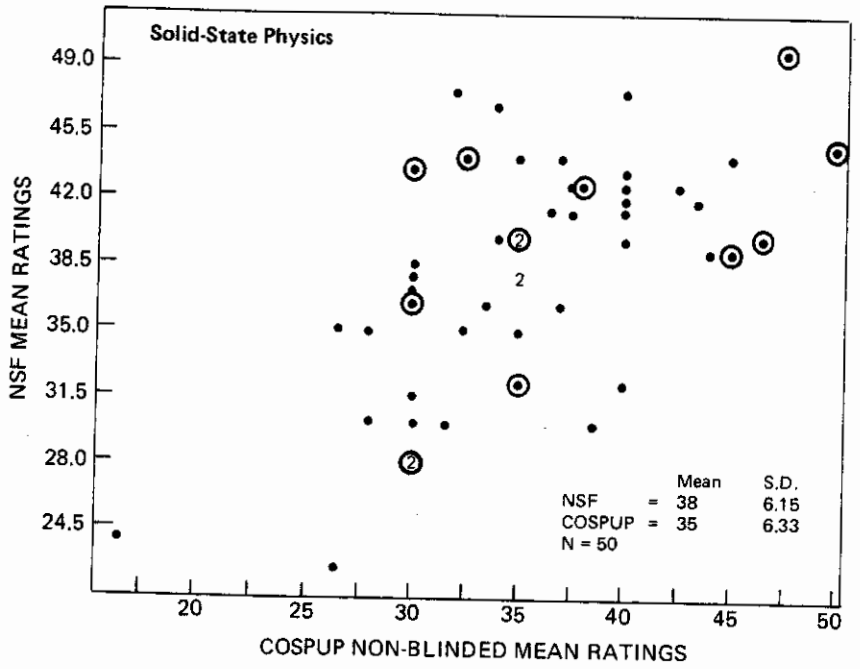

FIGURE 3 Scattergram of mean NSF ratings with mean COSPUP ratings of non-blinded individual proposals. Encircled dots indicate proposals that received two or less reviews from either COSPUP or NSF reviewers.

(best) quintile containing proposals ranked 1, 2, ..., 10. When there were ties for proposals at the borderline, they were apportioned among the quintiles. Thus, if proposals were tied at rank positions 9, 10, 11, 12, 13, each proposal is weighted as .4 proposal in the first quintile and .6 proposal in the second quintile. Also, NSF ties at the center value of 25.5 were apportioned. Hence, fractional reversals may be reported.

For those unfamiliar with statistics, we pause to point out that if a decision on the 50 proposals were made by flipping a coin, in a purely random way, we would obtain a 50 percent reversal rate on average. That is, each NSF decision would have approximately a 50 percent chance of being reversed by the coin flip. On the other hand, if the COSPUP reviewers were to have rated the 50 proposals in such a way that there was complete agreement with the NSF reviewers on which proposals fell into the top 25 and which fell into the bottom 25, the reversal

TABLE 7 Reversals in Three Programs: Comparing NSF Ratings with COSPUP Non-Blinded

Quintile Based Upon COSPUP Non-Blind Rating	% Reversals							
	Chemical Dynamics		Economics		Solid-State Physics		Three Fields Combined	
	NSF Mean Rating	NSF Decison	NSF Mean Rating	NSF Decision	NSF Mean Rating	NSF Decision	NSF Mean Rating	NSF Decision
1	26	26	20	5	23	16	23	16
2	24	24	45	45	22	24	30	31
3	60	60	30	28	49	43	46	44
4	20	20	45	42	34	29	33	30
5	20	20	0	0	6	11	9	10
Overall	30	30	28	24	27	25	29	26
N Cases	50	50	50	50	50	50	150	150

rate would be zero. Thus, we would expect to find a reversal rate somewhere between zero and 50 percent; a rate above 50 would require a negative correlation between the two sets of rankings.

What we found is summarized in Table 7. When the three fields are combined in the first COSPUP quintile, 23 percent of the proposals fall in the bottom half of the NSF order. Nine percent of proposals in quintile 5 fell in the top half of NSF ratings. Reversals within quintile 3 are, of course, of two kinds. In this range of borderline decisions, a reversal rate as high as 50 percent would not be surprising. Percentages within quintiles for each field are based on only 10 proposals and are therefore influenced by large sampling errors.

The three programs show similar overall reversal rates. They turned out to be between 24 percent and 30 percent, computed in each of the two different ways. If the borderline group (quintile 3) is omitted, every figure is between 20 percent and 28 percent.

Later in this chapter, after we have discussed disagreements among reviewers of a proposal, we shall return to these reversal rates and their significance.

Components-of-Variance Model

The variance in reviewer ratings that results in reversals has different sources, and it is the purpose of this section to sort them out. Because the analysis is technical, we will begin by briefly describing in lay language the sources of variability we focus on. Then we will present the statistical method and its results. Finally, we will restate the results in nontechnical terms.

First, reviewers' responses to proposals differ because proposals differ in quality. That is easily dealt with statistically by taking as a rough indicator of "quality" of a proposal the mean of all its ratings by both NSF and COSPUP reviewers. This leads to a measure of the variation in quality of proposals (σ_p^2 below) that can be compared with other sources of variation.

Second, the NSF review procedures and the COSPUP procedures were not identical. On the average, there may be systematic differences between NSF reviewer responses to all proposals and COSPUP reviewer responses. Earlier in this section we noted this "method effect" when we pointed out that COSPUP reviewers were on the average

slightly harsher than NSF reviewers. Since funding decisions are based on rankings, this "methods effect" is not important (but we did not ignore it in the mathematical analysis).

Even after compensating for the average "methods effect," reviewers may disagree in their ratings of a proposal because they are members of two groups selected differently--NSF reviewers as opposed to COSPUP reviewers. This "interaction effect" (σ_I^2 below) between proposals and evaluation method is important. If it is large, then the NSF reviewer group and the COSPUP reviewer group evaluated proposals differently. If it is small, they did not and we would not be able to detect any bias in the selection of NSF reviewers.

Finally, what is left are effects not already discussed. Their variation is denoted by $\sigma_{R,i}^2$ below and measures the reviewer variation within a given evaluation method, i. Later we speculate on reasons for such variations.

The statistical method for analyzing these variances is called analysis of variance. The standard method is not applicable because the assumptions under which it is derived are not valid here. A special model was designed for us by Lee Cronbach and Jack Kiefer and the computer calculations were made by Gary Simon. We are grateful for their help.

In this model we represent the rating, Y_{ijk}, given by the k^{th} reviewer under method i on proposal j by

$$Y_{ijk} = \alpha_i + b_j + c_{ij} + e_{ijk}.$$

The interpretation of the individual terms is as follows:

α_i is the effect associated with evaluation method i.

b_j is the effect associated with proposal j.

c_{ij} measures the extent to which the rating on proposal j depends on the evaluation method i.

e_{ijk} is the residual effect not assignable to the preceding factors; it includes both systematic and unsystematic effects entering the ratings of proposal j by the k^{th} reviewer in the i^{th} set. "Method" here refers to NSF reviewers, COSPUP reviewers of non-blinded proposals, and COSPUP reviewers of blinded proposals;*

*Denoted henceforth by COSPUP (NB) and COSPUP (B) respectively.

for the sake of contrast, two methods were analyzed at a time.

We consider the α_i to be fixed effects and the others to be independent random variables with mean zero. We can decompose the variance associated with proposals under method i into three components:

$$\mathrm{Var}(Y_{ijk}) = \sigma_p^2 + \sigma_I^2 + \sigma_{R,i}^2,$$

where

σ_p^2 reflects the intrinsic variability of the proposals, i.e., the variance of b_j (assumed to be independent of i and k);

σ_I^2 represents the interaction of proposals and evaluation method, i.e., the variance of c_{ij} (assumed to be independent of k); and

$\sigma_{R,i}^2$ is the variance of e_{ijk}, i.e., the reviewer variance within method i (assumed to be independent of j and k).

Before giving the results, we explain their interpretation, using one sub-analysis as example. The analysis of ratings of proposals in Solid-State Physics by NSF and COSPUP (NB) produced five numbers, which may be grouped as follows:

Variance among proposals*

$\hat{\sigma}_p^2$ 24.43

Variance associated with different effect of rating process on different proposals:

Interaction of method and proposal $\hat{\sigma}_I^2$ 0.17

Variance among reviewers:

Method 1 (NSF): $\hat{\sigma}_{R,i}^2$ 48.93

Method 2 COSPUP (NB): $\hat{\sigma}_{R,2}^2$ 50.24

Constant difference between methods:

$\alpha_1 - \alpha_2$ 2.7.

*(^) denotes the estimate of (^), obtained from our data.

The last number is not directly comparable to the others, as the difference is expressed in the scale units and the variances in squared units.

Decisions are based on average ratings, not single ratings. For purposes of interpretation it is convenient to assume that n ratings are obtained on each proposal. If this process were repeated with many sets of such reviewers, the variance "over reviews" (i.e., over average ratings for the same proposal) depends on the number of reviewers in the set. This dependency is taken into account if we write 48.93/n for the reviewer variance within Method 1 and 50.24/n for that within Method 2. The numbers above are variances for single reviewers (n = 1).

The "method effect" of 2.72 repeats what was said earlier: the mean of all NSF reviews differs by 2.72 points on the 50-point scale from the mean of all COSPUP non-blind reviews. The difference is of no practical consequence. Decisions derive from rankings, and ranks are not affected by this "constant difference." The variations that affect rankings appear in the second category. It will simplify the discussion to average the values for the two methods, so that we have:

Interaction	0.17
Reviewers	49.58/n

If n = 1, the total "error variance,"* the sum of the above two components, is then estimated to be 49.8; if n = 4, the estimate drops to 12.6.

Each variance is a squared standard deviation; standard deviations are more familiar than variances and in some ways easier to interpret. As a rule of thumb, one can say that values in a distribution rarely depart from the mean by more than two standard deviations. (Assuming a normal distribution, such departures occur about five times in 100 cases.) Taking the square root of each variance restates the information above so that the rule-of-thumb can be applied. We also divide the method effect by 2 to express it as a standard deviation.

*Defined as .17 + 49.58/n, reflecting both reviewer variations and differences in the effects of the methods.

Table 8 displays the amount of variation among ratings attributable to various sources (effect sizes) when the number of reviewers per proposal is one and also when it is four. To simplify the table, the "Reviewers" figures are based on the averages of the two varying components reported earlier for reviewers in methods 1 and 2. The term "Effect" means the proposal component of variance, the interaction component of variance, and the reviewer component of variance divided by n ($\hat{\sigma}^2_p$, $\hat{\sigma}^2_I$, and $(\hat{\sigma}^2_{R,1} + \hat{\sigma}^2_{R,2})/2n$, respectively).

For example, the proposal component of the total variance (σ^2_p) describes the variation to be expected among the b_j's, that is, among the "virtual" ratings that would be obtained by averaging a very large number of ratings of each proposal, half of them obtained under each method. According to the rule-of-thumb, most virtual ratings fall within 10 points of the mean for all proposals, that is, within a 20-point range on the 50-point scale. The error standard deviation of 7.1 indicates that the rating given by a single reviewer is likely to be within about 14 points of the virtual rating. Averaging four ratings reduces this spread to plus or minus 7 points; similar calculations can be made for other values of n. The standard deviation derived from $\hat{\sigma}^2_I$ is small; although some proposals fare better under one method than another, these variations are unlikely to move the proposal more than one point away from the virtual rating. Clearly, nearly all the variance here arises from differences among reviewers and not from the method of selection.

Tables 9, 10, and 11 present all three analyses of Chemical Dynamics, Economics, and Solid-State Physics. Now, of course, the methods are kept distinct because we want to examine the contrast between the three methods, each pair in turn. The NSF/COSPUP (NB) contrast will be emphasized here; discussion of findings specific to COSPUP (B) reviews is left to the next section.

Results in Chemical Dynamics and Solid-State Physics are essentially similar, and the interpretation worked out above for one contrast in Solid-State Physics is applicable to all six contrasts in these fields. (The larger interaction component in three of the contrasts involving reviewing of blinded proposals has no important consequence.)

TABLE 8 Detailed Examination of Estimated Effect Sizes for One Data Set[a] (Solid-State Physics)

Effect	Effect size when n = 1		Effect size when n = 4	
	Variance or difference	Standard deviation	Variance or difference	Standard deviation
Proposals	24.4	4.9	24.4	4.9
Error	49.8	7.1	12.6	3.5
Interaction	0.2	0.4	0.2	0.4
Reviewers	49.6	7.0	12.4	3.5
Method	2.7	1.4	2.7	1.4

[a]The data set is NSF and COSPUP non-blinded.

In the NSF/COSPUP (NB) contrast the two sources of error (interaction and reviewer disagreement) imply an error standard deviation of 9.6 in Economics if only a single rater is used, and one of 4.8 if four raters are averaged. This implies that ratings of the same proposal obtained by the two methods from many teams of four reviewers would range as much as 10 points from the true ratings in either direction.

We conclude: (i) there was no important systematic difference between the NSF and COSPUP (NB) methods in any of the fields and (ii) reversals in the NSF/COSPUP (NB) contrast appear to arise almost wholly from reviewer disagreement (within methods).

The analysis above shows that nearly all the variability in reviewer responses arises from differences among reviewers and very little is due to the group they were in (NSF or COSPUP (NB)). That is, $\hat{\sigma}^2_I$ is small in comparison with the $\hat{\sigma}^2_{R,i}$. We can state that (i) we find no evidence of systematic bias in the selection of reviewers by NSF program directors, and (ii) in comparing NSF with COSPUP (NB), reversals in the funding decision as described earlier in this section are almost wholly due to reviewer disagreement within each method.

TABLE 9 Estimates of Effect Sizes for Proposals in Chemical Dynamics

Contrast	Size of effect[a]				
	Proposals	Interaction	Reviewers[b]		Method[b]
			1	2	
NSF with non-blind	23.67	1.18	55.91/n	56.67/n	2.73
NSF with blind	18.38	2.59	55.91/n	83.10/n	2.59
Blind with non-Blind	23.91	0.00[c]	83.10/n	56.67/n	.15
NSF with non-blind	4.9	1.0	7.5[d]	7.5[d]	1.4
NSF with blind	4.3	1.6	7.5[d]	9.1[d]	1.3
Blind with non-Blind	4.9	0.0	9.1[d]	7.5[d]	0.1

[a]Numbers in the upper half of the table are variances or (in the method column) differences between means. In the lower half of the table the numbers are standard deviations.

[b]The method listed first in the "Contrast" column is Method 1 in the "Reviewer" column. A positive entry in the "Method" column indicates that the average rating was higher in Method 1 than in Method 2.

[c]The calculated estimate was negative.

[d]The standard deviation describes the spread among single raters. For n other than 1, each value is to be divided by the square root of n. Thus with n = 4 each of these values would be halved.

TABLE 10 Estimates of Effect Sizes for Proposals in Economics

Contrast	Size of effect[a]				
	Proposals	Interaction	Reviewers[b]		Method[b]
			1	2	
NSF with Non-Blind	58.33	0.00[c]	89.22/n	96.25/n	2.14
NSF with Blind	52.75	3.76	89.22/n	100.49/n	4.85
Blind with Non-Blind	54.41	1.05	100.49/n	96.25/n	-2.71
NSF with Non-Blind	7.6	0.0	9.4[d]	9.8[d]	1.1
NSF with Blind	7.3	1.9	9.4[d]	10.0[d]	2.4
Blind with Non-Blind	7.4	1.0	10.3[d]	9.8[d]	-1.4

[a]Numbers in the upper half of the table are variances or (in the method column) differences between means. In the lower half of the table the numbers are standard deviations.

[b]The method listed first in the "Contrast" column is Method 1 in the "Reviewer" column. A positive entry in the "Method" column indicates that the average rating was higher in Method 1 than in Method 2.

[c]The calculated estimate was negative.

[d]The standard deviation describes the spread among single raters. For n other than 1, each value is to be divided by the square root of n. Thus with n = 4 each of these values would be halved.

TABLE 11 Estimates of Effect Sizes for Proposals in Solid-State Physics

Contrast	Proposals (Size of effect[a])	Interaction	Reviewers[b] 1	Reviewers[b] 2	Method[b]
NSF with Non-Blind	24.43	0.17	48.93/n	50.24/n	2.72
NSF with Blind	27.46	3.17	48.93/n	72.14/n	3.02
Blind with Non-Blind	28.11	2.90	72.14/n	50.24/n	-0.31
NSF with Non-Blind	4.9	0.4	7.0[c]	7.1[c]	1.4
NSF with Blind	5.2	1.8	7.0[c]	8.5[c]	1.5
Blind with Non-Blind	5.3	1.7	8.5[c]	7.1[c]	-0.2

[a]Numbers in the upper half of the table are variances or (in the method column) differences between means. In the lower half of the table the numbers are standard deviations.

[b]The method listed first in the "Contrast" column is Method 1 in the "Reviewer" column. A positive entry in the "Method" column indicates that the average rating was higher in Method 1 than in Method 2.

[c]The standard deviation describes the spread among single raters. For n other than 1, each value is to be divided by the square root of n. Thus with n = 4 each of these values would be halved.

Reviewer Disagreement

Since the disagreement among reviewers of the same proposals is the major cause of reversals, we now look at the reasons for these disagreements.

Reviewer differences are in fact a natural and desirable feature of a judging process. Reviewers differ in their convictions about the values of different lines of investigation and about the usefulness of different kinds of evidence. Moreover, one reviewer may detect a flaw in reasoning that another misses. It is for this reason that multiple opinions are solicited, and that the program director is authorized to consider qualitative comments in addition to the numerical ratings.

Moreover, an important quality that must be judged in a proposal is its originality and potential for advancing science by large steps. This quality involves going beyond known science and carries with it large uncertainties about the final significance of the proposed research. Such highly original work, breaking with past traditions, is bound to be evaluated differently by experts in the field, who often have different views of the most promising directions for future work.

Thus, reviewer disagreements are to be expected at the frontiers of research. In an attempt to find some pattern in them we selected proposals with significant differences between the NSF mean rating and the COSPUP non-blinded mean rating, differences that would have created reversals. Then Dr. Albert Clogston (Solid-State Physics) and Dr. William Spindel (Chemical Dynamics), with the aid of some of their colleagues, examined the reviews. Both observed that there was a wide variety of reasons for disagreement and concluded that no summary statement would be meaningful. Clogston added, however, that differing research interests of the reviewers were apparent in the four cases he examined. He wrote:

> The analysis of these four cases leaves me with the impression that three of the proposals I analyzed received higher mean scores from the NSF peer reviewers than from the COSPUP(NB) reviewers because the NSF reviewers were in each case predisposed to support the particular line of research involved. Another proposal which received a low mean score from the NSF reviewers appears to have suffered because the proposal was

interdisciplinary in nature and was reviewed by members of two different constituencies who did not share a commitment to both aspects of the proposed research. These four proposals are a small sample of the fifty Solid-State Physics proposals included in the Phase II study of peer review in the NSF and do not warrant general conclusions. It seems to me, however, that there is some evidence in these cases that peer reviews can be influenced by the predisposition of the reviewer toward the importance of a given line of research in addition to his judgment about scientific merit. If this is the case, and if NSF program directors, in the interest of obtaining well-informed reviews, consistently select peer reviewers who are closely associated with the line of research involved in a given proposal, the result may mean scores compressed into a higher range than would be assigned by reviewers with a broader range of backgrounds. It seems to me the effect of this would be to make the decision for funding of marginal projects less reliable. It also seems possible that interdisciplinary proposals would suffer from such a practice. I think we should look for this effect by comparing the distribution of mean scores of NSF reviewers with the distribution of mean scores of COSPUP(NB) reviewers.

This exercise raises once again in my mind the question as to what peer review actually means as practiced by the NSF. Does it mean review by a group of peers in a broad field of science such as solid state physics (or chemical dynamics or economics), or does it mean review by members of a much smaller group of people expert in a particular line of research?

Another possible reason for reviewer disagreement, of course, is that scientists, like all others, have personal preferences and biases.* No doubt reviews of

*Two scientists may have exactly the same opinion of a proposal and yet rate it differently because the adjectival ratings have different meanings for them.

some proposals have been influenced by personal attitudes toward the applicant. Statistical methods cannot discern exceptional cases of bias. But the fact that such personal feelings are not a primary cause of disagreement is indicated by the analysis of the blinded reviews in the next section. Though many reviewers did not know the identity of the author of a proposal, we found substantially the same level of reviewer disagreement.

The level of reviewer disagreement suggests that between 25 and 30 percent of NSF funding decisions could be reversed were the ratings made by another equally qualified group of reviewers. Some of us are not surprised by this finding. We believe such differences of opinion are healthy and make for openness of the system to proposals that challenge accepted norms of scientific thought. It is important to recognize that the uncertainty involved is intrinsic to the nature of the highest quality research and does not represent an inability or failure of peer review to identify the most worthy proposals for funding. Furthermore, quite a number of reversals are inevitable when there are many highly qualified proposals and limited funding makes close calls necessary.

Some of us found the percentage of reversal higher than we expected. In the middle quintile it is not surprising, but in the top quintile it is. The findings mean this for the ten proposals rated at the top by NSF: If another set of qualified reviewers were chosen to evaluate these proposals, seven or eight of the ten would be funded. If a coin were tossed instead to determine funding for all 50 proposals, on the average five of the ten would be funded.

We won't pretend to guess the many responses to the previous paragraph. Many will say, given the difficulty of judging proposals at the frontiers of science, that peer review does very well indeed. others will say this complicated system does not buy you much. But one must remember that the coin-tossing alternative is predicated on maintaining the same quality of proposals. Although the evidence is fragmentary, it seems likely that there is a large element of self-selection among the applicants to NSF. They generally have active research programs and superior past research performance, or have been recently trained by scientists who do have them. Were a less stringent review procedure adopted, many more less-qualified proposals would be received. Suppose, for example, 50 more were received in a discipline and still

only 25 were to be funded. Then the coin-tossing method would select only two or three of the ten most talented scientists for funding.

The percentage of reversals is also directly tied to the stringency of funding. If more funds were available, NSF would receive more applications, and the self-selection hypothesis implies that the new applications would not be as qualified as the present ones. Consequently the reviewer variation over all proposals would increase in proportion to that on the same proposals, thus decreasing the percentage of reversals.

In its recommendations elsewhere in this document, COSPUP has suggested some changes in the present peer-review mechanisms prompted by the level of reviewer disagreement. Certainly we need to understand the causes of reviewer disagreement better in order to cope with it better. In any case, the level of disagreement at the extremes of the sample is a major finding of this study.

Correlations of Ratings with Status Characteristics of Applicants

The components-of-variance analysis showed only small differences (i.e., the interaction variance was relatively low) in the ways in which NSF and COSPUP non-blinded reviewers evaluated the set of proposals. Nevertheless we asked whether the differences found in mean ratings were correlated with the characteristics of applicants. If NSF reviewers were being selected to produce more favorable reviews of proposals submitted by eminent scientists, then the NSF reviews should correlate with the eminence of the applicants more closely than do the reviews obtained from our experimental group of raters.* Table 5 shows that this is not the case. We again find no evidence here for bias in favor of eminent, established scientists.

*We remind the reader that COSPUP reviewers were chosen by selectors given blinded copies of proposals and only about 15 percent indicated they knew who the principal investigator was. So COSPUP reviewers were not chosen to be easier on eminent investigators.

Section 3

RESULTS RELATED TO BLINDED PROPOSALS

A second purpose of this study was to assess the feasibility and effect of reviewing blinded NSF proposals, as mandated by Congress. We report in this section the results of our experiment with blind reviewing. We discuss the extent to which blinding proposals was possible and the difficulties in reviewing blinded proposals. We then compare NSF ratings with COSPUP(B), i.e., COSPUP ratings of blinded proposals. Finally we come to reversals and the components-of-variance analysis.

The Possibility of Blinding

The difficulty of altering a proposal so that a qualified reviewer will not be able to guess the author has already been discussed (pp. 17-19). We asked the reviewers of blinded proposals to guess the identity of principal investigators. Approximately half of the reviewers in Chemical Dynamics and Solid-State Physics made the attempt, while in Economics only 37 percent did. Table 12 presents the results.

In Chemical Dynamics, 50 percent made the attempt to identify the principal investigators. Among the attempted identifications, 80 percent were correct, demonstrating that many proposals in Chemical Dynamics could not be blinded with the method we used. Columns 3 through 6 of Table 12 give the corresponding results for Economics and Solid-State Physics. Note that 93 percent of the identifications offered in Economics were correct. In some cases, incorrect identifications included past collaborators of the actual principal investigators, or persons doing very similar work.

TABLE 12 The Proposals of Reviewers Receiving Blinded Proposals Who Guessed at the Identity of the Principal Investigator(s)

	Chemical Dynamics		Economics		Solid-State Physics	
	N	(%)	N	(%)	N	(%)
No Guess Made	105	(50)	125	(63)	100	(49)
Correct Guess	86	(40)	69	(35)	87	(43)
Incorrect Guess	21	(10)	5	(2)	15	(7)
Missing Data	0		0		1	
Totals	212	(100)	199	(100)	203	(99)

These results suggest that blinding cannot be expected to conceal the identity of the investigator. This should be no surprise, for one criterion for selecting referees is their extensive knowledge of a research specialty. These experts know the work of other active researchers in their specialties. In some small research areas all the qualified reviewers will know the work of almost every applicant. So, unless the proposal is altered to the point of losing essential content, it may be impossible to conceal the identity of the author.

In fact, for only a very few proposals could we obtain reviews from reviewers without some awareness of the proposal's authorship; with the type of blinding attempted here, it is difficult to get reviews of most proposals by reviewers not guessing the identities of the principal investigators.

We asked whether the characteristics of reviewers had anything to do with whether they could guess the identity of the principal investigator. The only information we had on reviewers was the prestige ranks of the departments in which academic reviewers were employed. We found no systematic relationship of this variable to the likelihood of guessing the identity of the principal investigator.

The next question concerns the characteristics of investigators. Was there greater success in blinding proposals when the publication record of the author was limited or when the author was in a less central or less distinguished institution? We correlated the proportion of total reviewers who did not identify authors with the

applicant characteristics used in Phase I. Specifically we used the proportion of blinded reviewers who did not try to guess the identity of principal investigators as the dependent variable and the characteristics of the principal investigators as the independent variables. Table 13 correlates the proportion of blinded reviews on which no guess was made with applicant characteristics (and also with NSF mean ratings, the mean ratings of the blinded proposals, and the NSF decision).

Many correlations are small to moderate, but almost all are negative, implying more identification for more highly rated proposals and more established applicants. Scientists whose research is relatively heavily cited, for example, are more likely to produce proposals that cannot be fully blinded.

Note that the patterns of these correlations are not wholly consistent across the three programs. For instance, attempts to guess the author's identity were more frequent for professionally older applicants in Chemical Dynamics and for professionally younger applicants in the other two programs. Similarly, while guesses were more frequent for Economics applicants who received their Ph.D.'s from relatively distinguished departments, the opposite was true in Chemical Dynamics and Solid-State Physics. With samples of 50, much of this fluctuation is attributable to change.

Finally, we find that average ratings of proposals tended to be lower when reviewers did not try to guess the author. For example, the correlation between the proportion of reviews on each proposal for which no guess was made and NSF mean rating was -.49 for Chemical Dynamics, -.37 for Economics, and -.44 for Solid-State Physics.

Difficulty in Reviewing Blinded Proposals

On the evaluation form sent the reviewers of blinded proposals we included two questions about the difficulty of reviewing blinded proposals. The data are presented in Table 14.

A large fraction of reviewers said a proposal was harder to understand with references removed. Over half said that a blinded proposal was harder to review. Apparently reviewers find blinded proposals more difficult to evaluate than non-blinded proposals. This did not affect their response rate, as we noted earlier.

TABLE 13 Proportion of Reviews on Each Proposal for Which No Guess Was Made Correlated with Other Variables

Independent	Chemical Dynamics (N=50)	Economics (N=50)	Solid-State Physics (N=50)
Professional Age	-.21	.39	.12
Rank of Ph.D. Department	-.23	.33	-.49
Rank of Current Department	-.32	-.59	-.21
Academic Rank	-.30	.12	.15
Log of Papers	-.32	-.26	-.26
Log of New Citations	-.32	-.36	-.31
Log of Old Citations	-.40	.04	-.15
Past Funding History	-.49	-.34	-.22
NSF Decision	-.45	-.43	-.55
NSF Mean Rating	-.49	-.37	-.44
Mean Blinded Rating	-.55	-.62	-.55

Comparison of NSF Ratings and Ratings of COSPUP Blinded Proposals

We now turn to the ratings by COSPUP (B) reviewers and compare them with ratings by both COSPUP (NB) and NSF reviewers. Table 15 shows a moderate correlation between COSPUP (B) and NSF, which is somewhat smaller than the correlation between COSPUP (NB) and NSF. (See scattergrams in Figures A.1-A.3 in Appendix A. Again Economics has the most encircled dots; see comment on p. 25.)

Reversals

We again consider reversals in a manner similar to that described in Section 2 (see pp. 29ff). The data are presented in Table 16.

Reversal rates are similar to those reported for COSPUP non-blinded reviews. The overall proportion of

TABLE 14 Questions on the Difficulty of Evaluating Blinded Proposals

	Chemical Dynamics	Economics	Solid-State Physics
Question 1			
Has the removal of references from this proposal made it difficult to understand?			
% saying yes	46% (211)	39% (196)	46% (203)
Question 2			
Is it more or less difficult to review a proposal on which an attempt has been made to conceal the identity of the principal investigator?			
More difficult	58%	59%	62%
Less difficult	1	2	1
Does not influence the difficulty	41	39	36
	100% (208)	100% (196)	99% (203)

reversals was slightly larger than before, as would be expected from the correlations reported above.

The overall percentage of reversals is 33 (compared with 27 percent before). When the COSPUP (B) reviews were compared with the NSF funding decisions, the reversal rate was 33 percent (compared with 25 percent previously).

In Chemical Dynamics and Economics, the reversal rate was highest in the middle (third) quintile, and lowest in the first and fifth quintiles. In Solid-State Physics, the highest reversal rates were in the second and fifth quintiles, with approximately equal reversal rates in the first, third, and fourth quintiles.

One should expect higher reversal rates when blinded reviews are compared with NSF reviews. COSPUP (NB)

TABLE 15 Correlations of Ratings by Three Sets of Reviewers

Correlation Between	Chemical Dynamics	Economics	Solid-State Physics
COSPUP (B) and NSF	.42	.55	.53
COSPUP (B) and COSPUP (NB)	.58	.63	.53
COSPUP (NB) and NSF	.60	.66	.62

reviewers were asked to use the same criteria as NSF reviewers, but COSPUP (B) reviewers were asked to use a different set. They were instructed (see Appendix B, page 76) that, even if they knew who the principal investigator was, they should evaluate the proposal strictly on its content.

Components of Variance

We refer the reader to Section 2 (pp. 32ff) for terminology. The magnitudes of effects attributable to proposals, method (blind/non-blind contrast), interaction, and reviewers were presented in Tables 9, 10, and 11. The component for proposals is almost the same as in the analysis of the two non-blind ratings, as might be expected. The method components show a tiny but consistent tendency for ratings of blinded proposals to average lower than the ratings of the same proposals in their original form. Perhaps the surprising thing is that loss of information did not greatly reduce the perceived worth of the proposals on average. Reviewer disagreement regarding blinded proposals was a bit larger than when the proposals were judged in their original form. For COSPUP reviewers in Solid-State Physics, for example, the standard deviation representing this disagreement rose from 7.1 to 8.5. Although the differences are small, they imply that more reviewers would be needed to achieve a given level of stability of average ratings for blinded proposals than to achieve the same level of stability with non-blinded proposals.

The estimates of the interaction effect differ from field to field, and, within a field, the estimate differs from one contrast to the next. We direct attention to

TABLE 16 Reversals in Three Programs: Comparing NSF Ratings with COSPUP Blinded

	% Reversals							
	Chemical Dynamics		Economics		Solid-State Physics		Three Fields Combined	
Quintile Based on COSPUP Blind Rating	NSF Mean Rating	NSF Decision	NSF Mean Rating	NSF Decision	NSF Mean Rating	NSF Decision	NSF Mean Rating	NSF Decision
1	30	30	15	10	20	20	22	20
2	40	50	53	46	38	38	44	45
3	52	52	62	64	22	22	45	46
4	30	30	45	50	20	30	32	37
5	10	20	25	10	40	30	25	20
Overall	32	36	40	36	28	28	33	33
N Cases	50	50	50	50	50	50	50	50

the contrast between the two COSPUP ratings (last line of each table). In these two sets of ratings, reviewer selection was controlled in the same manner and the interaction can therefore be strictly attributed to blinding and not to reviewer differences. The standard deviations representing this interaction effect are 0.0, 1.0, and 1.7 in Chemical Dynamics, Economics, and Solid-State Physics, respectively. We conclude that the systematic effect of blinding is small. That is, if a large number of blinded ratings were obtained for each proposal, few proposals would rank appreciably above or below their standing in a non-blinded competition with many reviewers. Reviewer disagreement, not blinding itself, accounts for nearly all the reversals in a blind/non-blind comparison.

Conclusions Concerning Blinded Proposals

We found: (i) As a practical matter, it was extremely difficult to conceal a proposal's authorship. (ii) More established applicants were more likely to produce proposals that cannot be fully blinded. (iii) Average ratings of proposals tended to be lower when reviewers did not try to guess the author. (iv) Reviewers found blinded proposals more difficult to evaluate than non-blinded proposals. (v) Reversal rates were similar but slightly higher than those for NSF/COSPUP (NB). (vi) More reviewers would be needed to achieve a given level of stability of average ratings for blinded proposals than to achieve the same level of stability with non-blinded proposals. (vii) The systematic effect of blinding is small. Few proposals would rank appreciably above or below their standing in a non-blinded competition with many reviewers. (viii) Reviewer disagreement, not blinding itself, accounts for nearly all the reversals in a blind/non-blind comparison.

Section 4

PEER REVIEW IN THE NATIONAL SCIENCE FOUNDATION: OBSERVATIONS OF THE COMMITTEE ON SCIENCE AND PUBLIC POLICY

Federal support for basic research in science has undergone sustained growth in the years since World War II. In part, this must reflect a general belief that the pursuit of fundamental knowledge is in its own right rewarding to any society. In part also, it is widely recognized that advances in science and advances in technology are to some extent coupled, and that this country's economic strength and social well-being are influenced by its technological position. Central among the many agencies with a part to play in the support of basic scientific research in many fields is the National Science Foundation (NSF). Hence any inquiry into the Foundation's exercise of its responsibilities touches on questions of national importance.

This study specifically addresses only the NSF's procedures for competitively evaluating proposals for basic research projects, the so-called peer-review system. With a few clearly specified exceptions, including certain categories of conferences and international travel, decisions on whether or not to fund all formal proposals depend primarily on external evaluations. About 80 percent of the NSF's total (fiscal 1981) budget of $994 million is assigned to programs of basic-research support that are subject to peer review. In addition, NSF performs a number of evaluative and advisory functions with regard to science policy that have no direct bearing on its basic research programs. It is also responsible for fiscal monitoring of on-going and completed projects and programs of policy-supportive or targeted research to which peer review is less applicable.

The maintenance of a fair and effective peer-review system is crucial to the NSF's purpose and mode of

operation. Yet peer review is unusual in government for it widely disperses responsibilities for the allocation of public funds. Peer review provides a kind of dialogue between the NSF and the scientific community concerning what science should be supported, and it has also come to symbolize for many scientists the orderly and critical debate that is necessary among qualified specialists if there are to be reliable judgments of scientific merit. A high proportion of this nation's productive scientists are involved from time to time in this debate. Their broad involvement, while surely not guaranteeing infallible judgments or ideal outcomes in all cases, contributes to a surer recognition of originality and promise among a very large and diverse number of scientific proposals.

Most members of the Committee on Science and Public Policy (COSPUP) have been involved in the peer-review process on numerous occasions. Concurrently with its sponsorship of the detailed investigations that were reported on in the Phase I study and in the earlier sections of this report, the Committee has examined this process repeatedly and at length over a five-year period. Its discussions have been informed by reports of preliminary findings from the investigations, by data supplied upon request by NSF representatives, and by the Committee's own substantial experience and familiarity with the process. Fairly broad areas of consensus have gradually become apparent, having to do both with the nature of the peer-review process and with possible improvements. In this final section we undertake to outline the Committee's own position, noting the points at which it rests directly on our investigative findings but not being limited to those.

(1) Proposals for basic research submitted to the NSF have two qualities that must be judged by peer review in deciding upon a fair and wise commitment of federal funds. The first quality is intrinsic merit and worthiness for funding in the light of what is already known about the field of science concerned, and involves such elements as the identification of important questions to be answered and sound and innovative experimental design. Though there will always be some spread in judgments about intrinsic merit by experts in the field, with regard to most proposals a consensus should be evident. There is plentiful evidence on which to base most such judgments, although, of course, experts may differ in their views of past work or proposed design.

The second and very important quality that must be judged is originality and potential for advancing science by large steps. This quality involves an extrapolation beyond known science and carries with it larger uncertainty about the final significance of the proposed research for the progress of science and technology. This is all the more true for the most original work, whose importance may become clear only after a period of many years. In spite of the greater uncertainty, it is obvious that a good number of highly original proposals must be funded if U.S. science is to retain world leadership. The challenge of peer review is to recognize high quality in spite of the uncertainty.

Highly original work breaking with past traditions is bound to be evaluated differently by experts in the field, who often have different views of the most promising directions for future work. This spread of views will be greater than the spread in their judgments of general merit because of the many unknowns. Such differences of opinion are healthy and make for openness of the system to proposals that challenge accepted norms of scientific thought. It is important to recognize that the uncertainty is intrinsic to the nature of the highest-quality research and does not represent an inability or failure of peer review to identify the proposals most worthy of funding.

(2) Peer review as a system is concerned with recognizing these basic qualities of science. But in an administrative sense it may be viewed as having two other primary requirements. The first is to obtain the soundest possible judgments as to which research undertakings will procure the greatest scientific advance for a given investment of resources. The second is to ensure equity for competing proposals, that is, the elimination of personal, institutional, and other considerations that are not strictly scientific. In very large measure these two objectives are mutually supportive; one will not be attained without a lively concern for the other. Yet they are not identical, or even conducive to the same specific judgments about proposals under all circumstances. The thrust of the concern for fairness is toward ranking only according to quantifiable criteria, and the elimination of subjectivity may argue for a high degree of anonymity for both proposers and referees. The referees' ability to recognize excellence, on the other hand, may be impeded if they are not able to take into account the previous

successes or "track records" of the proposers. Thus the challenge facing the NSF is to find a balance between these two goals, not invoking a rigid formula but retaining the flexibility that is required by the diversity of scientific activity. It is the primary consensus of the members of COSPUP that the system of peer review the NSF presently employees is a creditable response to this challenge. It is both

(a) effective in the selection of proposals with superior likelihood of leading to significant scientific advances, and

(b) fair in its concentration on the scientific attributes of proposals without regard to the personal contacts and institutional affiliations of the investigators making them.

(3) As indicated in the Introduction to this report, the issue immediately prompting the Phase II study is the feasibility of concealing the identity of proposers from reviewers. "Blind" reviewing has the appearance of fairness, but it is our unanimous judgment that "blinding" of proposals is both impractical and ill advised. First, authorship of a scientific proposal is hard to hide from knowledgeable reviewers. Even after fairly thorough deletion of references and descriptions of available facilities and previous work, some two fifths of the reviewers in our experiment were able to identify the investigators (see Table 12). Second, more extensive "blinding" was found to produce alterations in proposal contents that severely compromised the intent, methodology, and quality of the proposed research (pp. 17-18). Third, reviewers reported greater difficulty in evaluating proposals whose authors they had not been able to identify, and were more frequently and widely in disagreement where this was the case (pp. 48ff). The general imposition of a "blinding" requirement therefore would require the involvement of large numbers of additional reviewers to permit the system to function with comparable effectiveness. Fourth, it is interesting to note that the proposals of investigators with numerous citations of their work have less chance of being "blinded" effectively, while proposals that were "blinded" obtained lower average ratings (pp. 46ff). This suggests that attempted "blinding" might tend to favor the most eminent, individually recognizable investigators over less prestigious applicants, which is exactly the opposite of the effect that is intended.

Most important, both Phases I (Section 3) and II (Draft pp. 25,44) of this study show that reviewers' comments and evaluations are, within the limits of our statistical study, essentially free from the "old boy" and other biases sometimes attributed to them. Attempted "blinding" would create many impediments to a fair and effective review process. To do this in order to correct for unsupported suppositions of bias could only weaken the competitive system already in place for the support of basic science.

(4) Some understanding of the subtlety and complexity of the NSF's task in selecting proposals for support is essential, both to evaluate the peer-review system's present performance and as a background to consideration of various proposals for making it more effective. COSPUP's appreciation for the difficulties and uncertainties inherent in attempting improvements has grown throughout the five years of this study. Principal among the difficulties with which any system of competitive evaluation must contend are the following:

(a) There is as yet little basis for comparing the research contributions and productivity of NSF applicants with those of the large field of other eligible scientists. The evidence at hand, however, suggests self-selection among the applicants (Phase I report, p. 21). It is likely, therefore, that a fairly large proportion of the proposals submitted to the NSF are of superior quality. This narrows the range of differences among them and exacerbates the problem of selection.

(b) As already indicated in the Phase I report (pp. 97-106), our study has strongly reinforced the general impression that evaluation of even the best science is not a matter of consensus and certainty. Deep-seated differences in approach and methodology are present in almost all fields, and evaluations can be expected to differ in many of the same ways research priorities differ among individual investigators. Differing evaluations of the same proposals by peer reviewers are therefore very common, including some cases in which a significant body of experts foresee high scientific accomplishment. It may even be that the most pioneering proposals often generate the

greatest diversity of opinions. We believe that there is no way the NSF can alter the peer-review process so as to avoid having to make decisions that are highly consequential for individual investigators on the basis of conflicting advice.

(c) A simple view of the evaluation process, probably rather close to the one held by most COSPUP members at the outset of this study, is that uncertainties are greatest near the cut-off line for NSF funding. This supposes something approaching a consensus, in other words, about the most and least promising proposals, but acknowledges that slight differences in evaluation of the middle range might tip the funding/no-funding decision either way. What our investigations now have shown, however, is that differences are more substantial than this view implies. Even excluding the middle quintile of NSF rankings adjoining the cut-off line for funding, for which reversals of decisions by different panels of reviewers at a rate approaching 50 percent would not be difficult to understand, we have now found reason to expect that such reversals would be fairly common. In our experiments, 20 to 28 percent of proposals receiving ratings above this quintile by one panel of reviewers received ratings below the cut-off level for funding from another panel selected according to roughly comparable standards (pp. 59-60). It is clear, in short, that evaluative differences are not confined to a limited group of proposals of seemingly marginal value. Many projects given the highest ratings by some groups of expert reviewers would receive ratings from other, similarly constituted groups that would be too low to permit funding.

(5) The NSF's present peer-review procedures differ in one respect from program to program, as outlined in the Phase I report (p. 7). Some rely on mail review alone and others supplement mail reviews with periodic meetings of panels of specialists drawn from the fields in question. It is not apparent in most cases how this procedural difference arose, but a connection with differences between scientific fields seems entirely plausible. Panels might be less effective, for example,

in fields characterized by a high degree of internal subspecialization. Almost certainly there are other variables as well. For example, more investigators are active in some fields. Some fields rely more than others on the NSF as a chief source of research support. Fields also differ in the extent to which they presuppose the sustained activity of relatively large and heavily instrumented research teams.

We doubt that there is any uniform formula of improvements in peer-review procedures capable of being applied with consistently positive effects across this range of conditions. Deliberate, carefully circumscribed testing of a number of possible changes is appropriate. In the past, the NSF has experimented on a small scale with new procedures thought to be applicable to particular programs or divisions. We warmly encourage a continuation of that approach.

(6) It is a major responsibility of NSF program directors to select peer reviewers whose qualifications permit an authoritative, balanced evaluation of each proposal. Ideally, in making these selections considerable knowledge of the substantive issues under consideration in individual fields will be coupled with a critical awareness of the caliber and direction of work of those active in them. The evidence available on the performance of the NSF in this respect is indirect and limited. It stems primarily from the experimental replication of review panels described earlier in this report, and that was made possible by COSPUP sponsorship and participation in the alternative reviewer-selection process. And it must be borne in mind that this replication involved only three NSF programs. But in spite of this qualification, similarities in the overall rates of acceptance and rejection between the original NSF reviewer panels and their COSPUP alternatives suggest that they were employing roughly similar standards of judgment as to proposal quality.

Also relevant on this question are interviews with program directors included by the Coles in their Phase I study (pp. 29-32). They support the impression that NSF program directors generally approach the selection of reviewers with full cognizance of the need to secure both expertise and balance. This impression, together with our failure to find statistical evidence for bias in the selection process, provides some assurance that adequately qualified reviewers are being selected. There is bound to be considerable variation in performance

among individual program directors, of course, in this as in other dimensions.

(7) A second major responsibility of program directors is for decisions on whether or not to fund proposals after they have been evaluated by peer reviewers. As in the NSF's pursuit of the twin goals of equity and scientific excellence, referred to earlier, two similar but not identical objectives must be kept in mind. On the one hand, the program director must act and be seen to act as the instrument of a more or less clearly articulated scientific consensus if the peer-review process is to have any meaning. On the other hand, peer reviewers are not unanimous in their advice on individual cases. Apart from differences over approach and methodology that have already been mentioned, evaluative comments and ratings are not all based on the same criteria. Some reviewers are closely familiar with the nuances of an investigator's work but unprepared to evaluate its potential contribution from the perspective of a wider body of knowledge. Others, less immediately informed, pay less attention to the specific approach proposed and base their judgments largely on what they perceive to be a hierarchy of importance and urgency among scientific problems. Such differences in viewpoint are often complementary even though they may diverge in their implications for particular decisions; beyond these, one must expect occasional interpretations and statements by reviewers that are contradictory among themselves.

For these reasons program directors cannot and ought not be completely bound by the numerical averages of the peer ratings whenever divergent viewpoints or criteria of evaluation become evident. Program directors do, for the most part, recognize the need for some latitude in decision making, as interviews with directors that were included in the Phase I report (Section 4) make clear. Yet it is somewhat disconcerting to note that the outcome in the great majority of cases nonetheless remains closely correlated with the numerical average of peer ratings. We believe that a reasonable exercise of discretion must be not only permitted but encouraged, within guidelines that good program directors will elicit from active researchers in the field. Comments and evaluations submitted by qualified peer reviewers should constitute the framework of the individual decisions they reach, but ought not uniformly be allowed to dictate the funding/no-funding outcome.

Such discretion should be exercised only for clearly articulated reasons. It is naturally open to later inquiry and criticism. Commendably, the NSF has already established procedures for reconsideration of proposals initially declined. These begin with the program director furnishing an explanation to the investigator at his or her request, including verbatim copies of the peer reviews (but without the reviewers' names or identifying data). There is also provision for subsequent reconsideration of the action by the assistant director and (upon request of the applicant institution) by the deputy director of the NSF.

Clearly, the power of discretion placed in the hands of the program director is considerable, even with the constraints of due process and later appeal. This only underlines the vital importance for the whole peer-review and proposal-selection process of securing program directors of the very highest quality.

(8) In using peer review, program directors would be wise by and large to direct greater attention to the numerous qualitative considerations raised by the reviewers in their written comments, moving away from reliance on single numerical ratings. Some criteria of judgment are shared by many reviewers. Hence it should be possible to ask reviewers to express separate judgments about several distinct attributes, replacing the single scale with a number of numerical scales. Among them we believe there might be separate rankings of each proposal for the following:

(a) Proposal quality, in relation to a small group of directly comparable proposals focused on the same immediate field of investigation.

(b) The expected scientific contribution from the immediate domain of investigation involved in a proposal to current or prospective advances in a whole discipline, such as chemical dynamics or mathematical economics. We favor a more conscious and systematic effort to distinguish between the qualities of a proposal relative to its immediate field and the broader promise that this field or line of investigation holds for the more inclusive discipline(s) of which it is a part. Both kinds of evaluation are valid and necessary parts of the review process. They would be more regularly available for program directors to consider if there were a means for

differentiating between reviewers' positions on each of them.

(c) Originality, or the extent to which it contains innovative elements, as differentiated from technical elaborations along lines that are already familiar. It is not at all certain how satisfactorily innovativeness can be judged. There may be substantial differences among fields in this respect. Although risks of failure go up with increasing innovation or unorthodoxy, some of the most important scientific advances can be expected to come from a small proportion of the seemingly most innovative or unorthodox proposals. Scientific advance depends to a disproportionate extent, after all, upon pioneering efforts in uncharted directions. Hence, ways must be found to identify such efforts for special consideration in spite of the greater uncertainties regarding their promise.

Several recent steps by the NSF reflect an increasing awareness of this responsibility. One is a provision for two-year extension of some existing three-year grants in which there have been especially creative research accomplishments, permitting the principal investigators to attack "higher-risk" research questions without the possible constraint of competitive review. Another, initially limited to the Chemistry Division, provides a different option to the repeated submission of full-scale proposals. Investigators are encouraged instead to submit the publications resulting from their current grants, together with very brief statements of the directions they propose to take next. While obviously attaching greater importance to the investigator's record of accomplishment, this second approach also loosens the restraint of detailed specification of future work. By so doing, it allows greater freedom to pursue unpredictable but highly promising research opportunities. These and similar experiments will need to be carefully monitored, but we warmly approve the scientific sensitivity and administrative openness that have led to them.

(d) Track record, or the quality of the scientific performance of the investigators(s) in the past, particularly during the immediately preceding period. The previous performance of the investigator(s) over the past five years or so is often relevant, for example, to judgments as to whether particular proposals are not merely promising in principle but likely to be successful in practice. Recognizing this, recently promulgated NSF guidelines place considerable stress on an investigator's recent record of performance. Final decisions generally rest more heavily on the quality of proposals themselves, to be sure, and emphasis on past work must be light or selective enough so that young investigators do not suffer because their records are short.

This is a matter requiring further experimental trial by the NSF. While separate numerical ratings might help to highlight and standardize some of these criteria of judgment, they might obscure others. We would not want more scales to be added if that were merely to lead to an even more complete dependence on numerical ratings. Reviewers might better be asked, for example, to discuss at greater length themes like those suggested in their written comments. In any case our wish is not to discourage but to supplement reviewers' responses regarding the features of proposals that appear most salient to them. It is important, first, that the criteria of judgment given special weight by reviewers be made more explicit if possible. Second, reviewers should be encouraged to offer evaluations based on multiple criteria with some degree of comparability among different reviews.

(9) Present NSF policy calls for a balance of criteria to be employed in the selection of reviewers. Some reviewers should be closely familiar with the abilities of the individual investigator, his or her institutional support system, and the methodologies and data in the immediate field of research. Others should be concerned with the effects of the proposed project on adjacent lines of scientific investigation, and sensitive to broader trends in the pertinent fields of scientific development. Those criteria often overlap, of course, and reviewers in any case may not play the parts for which they have been primarily chosen.

Thus we feel that the strategy for balancing "narrow" judgments as to a proposal's quality and "broader" consideration of its importance for a discipline might in some circumstances be extended into a sequential process. Evaluations of its details and of the quality of its specific contributions would first be sought from researchers concerned with the same immediate group of scientific procedures or problems. The investigator probably should be encouraged to propose the names of some of those on this first panel of reviewers. The program director then would briefly paraphrase the reviews of this first panel, and, if generally positive, send the proposal together with these comments to a different set of reviewers selected as being appropriate for the broader stage of evaluation.

This procedure runs the risk of introducing excessive delays in funding. Undue delay is indeed to be avoided. We welcome the NSF's declaration, which is responsive to this concern, of the policy of completing action (to the stage of declination, withdrawal, or recommendation of an award) within nine months after receipt of a proposal by the Foundation. Small-scale experimentation with sequential review in selected cases may nevertheless be justified, since it could be especially helpful in new, controversial, multidisciplinary or borderline fields. To state briefly the principle underlying the suggestion, we believe that the caliber of funding decisions will generally be improved if they need not be made as a single act in a compressed time period. An orderly sequence of steps is to be preferred, since it allows ample possibilities for feedback. But we acknowledge that any general prescription like this must have very generous allowance for exceptions.

(10) Advisory-panel reviews have value in many circumstances as a supplement to mail reviews; that is implicit in the principle just outlined. Most members of COSPUP make this positive assessment of them. Panels have the advantage of providing a means of evaluating uncertainties through discussion, of assuring a complementarity of "broad" and "narrow" perspectives in the summary judgment, and of permitting some differences of viewpoint to be resolved through the give-and-take of a meeting. They are probably more effective than mail reviews alone in guarding against over concentration of research on fashionable topics, since large numbers of proposals come under scrutiny at one time. Interaction with a well-chosen panel on a series of proposals also

can only improve the understanding of a program director as to where the priorities of his or her field lie.

On the other hand, for some, elitist or oligarchic distortions may be perceived as being introduced into the domain of science by the use of panels even if their membership is periodically rotated. Panels are also more costly than mail reviews alone, and this at a time when there are insufficient funds to support all projects of acceptable quality. In fields with marked internal specialization, inordinate weight may be given to the views of the single panel members most immediately familiar with the area of a given group of proposals. Hence we do not advocate the introduction and use of panels in all programs. Intermediate measures, such as conference calls arranged with a group of specialists to discuss several proposals falling in the same field, may obtain many of the benefits of panels while avoiding some of their drawbacks.

(11) Implicit in the principle of reaching funding decisions in an orderly stepwise fashion that was enunciated at the end of item 9 is a case for periodically reviewing the decisions made within each NSF program. Unfunded as well as funded proposals should be included in the field of examination. The NSF has, in fact, recently introduced an oversight function at three-year intervals as a requirement for all programs. Committees assembled for the purpose are assured full access to rejected as well as accepted proposal files and to other relevant documentation. Provision is made for committee review of a sample of proposals, although their major concern probably will be less with the merits of individual proposals than with aggregates and trends. Among the questions asked by such committees should be the following:

(a) Was what is now regarded as the best science recognized and supported earlier by a given program?

(b) Are there circumstances in which it is useful to compare the final outcomes of research carried on under certain proposals with the evaluations initially made of those proposals by reviewers? (We are not thinking here of a general program of post-performance evaluations but of special problems such as projects in new or multi-disciplinary areas.)

(c) Was the program adequately responsive to innovative thrusts, to the needs of young researchers, and to the somewhat different criteria that may obtain in multidisciplinary areas on the margins of a program?
(d) Viewing a program director's decisions over the long run, was the director sufficiently cognizant of the major research trends in the field and of knowledgeable individual scientists who could be turned to for sound evaluations?
(e) Were the program director's priorities generally consistent with those of the field?
(f) Is there more justification for dispersing the available funds widely in a particular field, or for concentrating them in order to provide greater support for the best individual investigators or teams?
(g) Apart from funding/no-funding decisions (the exclusive focus of this study), have approved proposals always been funded at levels that allow the investigation to go forward, or have some of those lying close to the lower margin of approval been awarded inadequate amounts with which to carry on sound experimental programs?

(12) It is a matter of importance that a sufficient number of scientists review each proposal, but we would be reluctant to have program directors make the numbers of reviewers the primary concern. Proposal reviewing can become cumulatively a heavy burden, especially for the active researchers whose appraisals are most valuable. An awareness of this problem presumably underlies the NSF's recent introduction of a 15-page limit on proposal length. We cannot visualize any substitute for the entirely voluntary system that is in use today, but it seems almost certain that under this system a concerted effort to secure additional numbers of reviews would lead to a declining response rate among reviewers. Moreover, it must be acknowledged that in some cases program directors are already encountering considerable difficulties in obtaining a necessary minimum of reviews for all proposals.

We conclude that the greatest need is not for a general increase in the number of peer reviewers but for assurance of a minimal number of reviewers for each proposal that will meet a standard of fairness. As a provisional number, we urge that the NSF take steps to

obtain no fewer than four reviews for each proposal, except where mail reviews are supplemented by review from some form of advisory panel. This may require seeking supplementary evaluations if the initial requests are not met with a sufficient number of responses within a reasonable time.

APPENDIX A

SCATTERGRAMS SHOWING DISTRIBUTIONS OF RATINGS
TABLES LISTING REVIEWER RATINGS OF ALL PROPOSALS
BY THREE METHODS

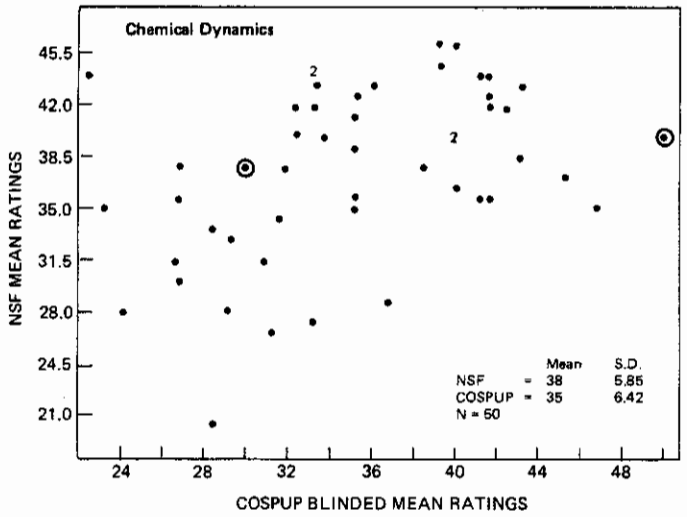

FIGURE A-1 Scattergram of mean NSF ratings with mean COSPUP ratings of blinded individual proposals. Encircled dots indicate proposals that receive two or less reviews from either COSPUP or NSF reviewers.

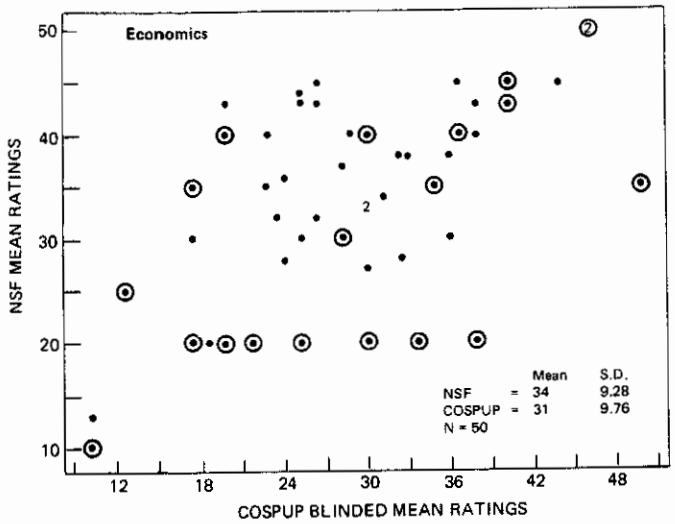

FIGURE A-2 Scattergram of mean NSF ratings with mean COSPUP ratings of blinded individual proposals. Encircled dots indicate proposals that received two or less reviews from either COSPUP or NSF reviewers.

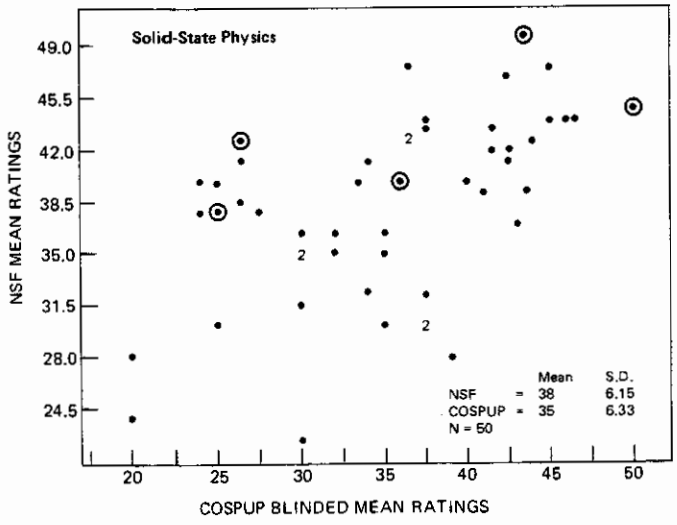

FIGURE A-3 Scattergram of mean NSF ratings with mean COSPUP ratings of blinded individual proposals. Encircled dots indicate proposals that received two or less reviews from either COSPUP or NSF reviewers.

TABLE A-1 Raw Data of Peer Reviews and Basic Statistics for the 50 Proposals in Chemical Dynamics: NSF Reviews

Proposal Number	Peer Review Score									Mean	Variance	Standard Deviation
	10	15	20	25	30	35	40	45	50			
101	–	–	–	–	1	–	3	–	–	37.50	25.00	5.00
102	–	–	–	–	1	–	–	1	2	43.75	89.58	9.47
103	–	–	–	–	–	–	4	–	–	40.00	0.00	0.00
104	–	–	–	–	1	–	3	–	1	40.00	50.00	7.07
105	–	–	–	–	1	–	2	–	3	43.33	66.67	8.17
106	–	–	–	–	–	–	2	–	1	43.33	33.33	5.77
107	–	–	–	–	–	–	3	–	2	44.00	30.00	5.48
108	–	–	1	–	–	–	3	–	1	38.00	120.00	10.95
109	–	–	–	–	–	–	3	–	1	42.50	25.00	5.00
110	–	–	–	–	–	–	3	1	–	41.25	6.25	2.50
111	–	–	–	–	–	1	2	–	2	43.00	45.00	6.71
112	–	–	–	–	–	–	4	–	1	42.00	20.00	4.47
113	–	–	–	–	–	1	1	–	3	45.00	50.00	7.07
114	–	–	–	–	–	1	3	–	3	43.57	39.29	6.27
115	–	–	–	–	1	–	3	–	1	40.00	50.00	7.07
116	–	–	–	–	1	–	2	–	2	42.00	70.00	8.37
117	–	–	–	–	1	1	3	–	3	41.88	56.70	7.53
118	–	–	–	–	1	–	1	–	3	44.00	80.00	8.94
119	–	–	–	–	–	–	2	–	3	46.00	30.00	5.48
120	–	–	1	–	–	1	1	3	–	38.33	96.67	9.83
121	–	–	–	–	–	–	3	–	2	44.00	30.00	5.48
122	–	–	–	–	–	1	4	–	–	39.00	5.00	2.24
123	–	–	–	–	–	–	3	–	2	44.00	30.00	5.48
124	–	–	–	–	–	–	4	–	–	40.00	0.00	0.00
125	–	–	–	–	–	–	2	–	3	46.00	30.00	5.48
126	–	–	–	–	–	–	5	–	1	41.67	16.67	4.08
127	–	–	2	–	–	–	1	–	–	26.67	133.33	11.55
128	–	–	1	–	3	–	–	–	–	27.50	25.00	5.00
129	–	–	–	–	3	–	1	–	1	36.00	80.00	8.94
130	–	–	2	–	1	–	1	–	2	35.00	190.00	13.78
131	–	–	–	–	1	1	3	–	–	37.00	20.00	4.47
132	–	–	–	–	1	1	1	1	–	37.50	41.67	6.46
133	1	–	–	–	3	–	1	–	–	28.00	120.00	10.95
134	–	–	–	–	2	1	1	–	–	33.75	22.92	4.79
135	–	–	–	–	2	–	1	1	–	36.25	56.25	7.50
136	1	–	–	–	1	–	2	–	1	34.00	230.00	15.17
137	–	–	2	–	1	–	–	1	1	33.00	195.00	13.96
138	–	–	–	–	5	–	1	–	–	31.67	16.67	4.08
139	–	–	–	–	2	–	2	–	–	35.00	33.33	5.77
140	1	–	1	–	1	–	–	–	–	20.00	100.00	10.00
141	–	–	–	–	2	1	3	–	1	37.86	48.81	6.99
142	–	–	–	1	2	–	1	–	–	31.25	39.58	6.29
143	–	–	–	–	1	2	1	–	2	40.00	70.00	8.37
144	–	–	–	–	2	–	3	–	–	36.00	30.00	5.48
145	–	–	–	–	2	–	3	–	–	36.00	30.00	5.48
146	–	–	–	–	2	–	3	–	–	36.00	30.00	5.48
147	–	–	–	–	2	–	2	–	–	35.00	33.33	5.77
148	–	–	–	1	3	–	–	–	–	28.75	6.25	2.50
149	–	–	–	–	4	–	–	–	–	30.00	0.00	0.00
150	–	–	2	–	2	–	1	–	–	28.00	70.00	8.37

TABLE A-2 Raw Data of Peer Reviews and Basic Statistics for the 50 Proposals in Chemical Dynamics: Reviews of Non-Blinded Proposals

Proposal Number	Peer Review Score									Mean	Variance	Standard Deviation
	10	15	20	25	30	35	40	45	50			
101	–	–	–	1	2	–	1	–	–	31.25	39.58	6.29
102	–	–	–	–	2	–	1	–	1	37.50	91.67	9.57
103	–	–	1	–	2	–	–	–	1	32.50	158.33	12.58
104	–	–	–	–	2	–	2	–	–	35.00	33.33	5.77
105	–	–	–	–	–	1	2	1	–	40.00	16.67	4.08
106	–	–	–	–	–	2	2	–	–	37.50	8.33	2.89
107	–	–	1	–	1	–	1	–	1	35.00	166.67	12.91
108	1	–	–	–	4	–	–	–	–	26.00	80.00	8.94
109	1	–	–	–	4	–	–	–	–	26.00	80.00	8.94
110	–	–	–	–	1	–	4	–	–	38.00	20.00	4.47
111	–	–	–	–	–	1	2	–	–	38.33	8.33	2.89
112	–	–	–	–	–	1	5	–	–	39.17	4.17	2.04
113	–	–	–	–	1	–	3	1	–	39.00	30.00	5.48
114	–	–	–	–	–	–	–	–	4	50.00	0.00	0.00
115	–	–	–	–	3	–	1	–	–	32.50	25.00	5.00
116	–	–	–	–	1	2	2	–	–	36.00	17.50	4.18
117	–	–	–	–	2	1	1	–	1	37.00	70.00	8.37
118	1	–	–	–	1	–	–	–	–	20.00	200.00	14.14
119	–	–	–	–	1	–	–	–	1	40.00	200.00	14.14
120	–	–	–	–	1	–	4	–	1	40.00	40.00	6.33
121	–	–	–	–	–	–	–	1	3	48.75	6.25	2.50
122	–	–	–	–	1	1	1	–	–	35.00	25.00	5.00
123	–	–	–	–	1	–	4	–	–	38.00	20.00	4.47
124	–	–	–	–	–	–	3	–	1	42.50	25.00	5.00
125	–	–	–	–	1	–	1	–	2	42.50	91.67	9.57
126	–	–	–	–	–	–	4	1	–	41.00	5.00	2.24
127	1	–	2	1	2	–	–	–	–	22.50	57.50	7.58
128	–	–	1	–	3	–	–	–	–	27.50	25.00	5.00
129	–	–	–	–	–	2	1	–	1	40.00	50.00	7.07
130	–	1	–	–	–	–	1	–	2	38.75	272.92	16.52
131	–	–	–	–	1	–	1	–	2	42.50	91.67	9.57
132	–	–	1	–	2	–	2	–	–	32.00	70.00	8.37
133	–	–	–	1	2	–	2	–	–	33.00	45.00	6.71
134	–	–	–	–	4	–	–	–	–	30.00	0.00	0.00
135	–	–	–	–	2	–	2	–	–	35.00	33.33	5.77
136	–	–	–	–	1	1	2	–	–	36.25	22.92	4.79
137	–	–	–	–	3	1	–	–	–	31.25	6.25	2.50
138	1	–	–	–	2	1	–	–	1	31.00	205.00	14.32
139	–	–	–	1	1	1	–	–	–	30.00	25.00	5.00
140	1	–	1	–	2	–	–	–	–	22.50	91.67	9.57
141	–	–	–	–	1	1	1	–	1	38.75	72.92	8.54
142	–	–	–	–	4	–	–	1	–	33.00	45.00	6.71
143	–	1	–	–	–	1	–	–	–	25.00	200.00	14.14
144	–	–	–	1	1	–	1	1	–	35.00	83.33	9.13
145	–	–	–	–	2	1	2	–	–	35.00	25.00	5.00
146	–	–	–	–	–	–	2	–	1	43.33	33.33	5.77
147	–	–	1	1	2	–	1	–	1	32.50	117.50	10.84
148	–	–	1	1	2	–	–	–	–	26.25	22.92	4.79
149	–	1	1	–	3	–	–	–	–	25.00	50.00	7.07
150	–	–	1	–	3	1	–	–	–	29.00	30.00	5.48

TABLE A-3 Raw Data of Peer Reviews and Basic Statistics for the 50 Proposals in Chemical Dynamics: Reviews of Blinded Proposals

Proposal Number	Peer Review Score									Mean	Variance	Standard Deviation
	10	15	20	25	30	35	40	45	50			
101	–	–	2	–	–	–	3	–	–	32.00	120.00	10.95
102	–	–	–	–	2	–	1	–	–	33.33	33.33	5.77
103	–	–	–	–	2	1	1	–	–	33.75	22.92	4.79
104	–	–	–	–	–	–	–	–	2	50.00	0.00	6.71
105	–	–	–	1	2	–	2	–	–	33.00	45.00	6.71
106	–	–	–	–	3	1	1	–	1	35.83	64.17	8.01
107	1	–	–	–	1	–	4	–	–	33.33	146.67	12.11
108	–	–	2	–	4	–	–	–	–	26.67	26.67	5.16
109	–	–	1	–	1	1	1	–	1	35.00	125.00	11.18
110	–	–	–	–	2	–	2	–	–	35.00	33.33	5.77
111	–	–	–	–	2	–	1	–	3	41.67	96.67	9.83
112	–	–	–	–	1	–	1	–	2	42.50	91.67	9.57
113	–	–	–	–	1	–	3	1	–	39.00	30.00	5.48
114	–	–	–	–	–	–	2	–	1	43.33	33.33	5.77
115	–	–	–	–	1	–	2	–	1	40.00	66.67	8.17
116	–	–	1	–	–	–	2	–	–	33.33	133.33	11.55
117	–	–	–	–	3	–	1	–	–	32.50	25.00	5.00
118	2	–	–	–	1	–	1	–	–	22.50	225.00	15.00
119	1	–	–	–	–	–	1	1	2	39.00	280.00	16.73
120	–	–	–	–	–	–	4	–	2	43.33	26.67	5.16
121	–	–	–	–	–	1	1	2	–	41.25	22.92	4.79
122	–	–	–	–	1	1	1	–	–	35.00	25.00	4.08
123	–	–	–	–	–	–	5	–	1	41.67	16.67	4.08
124	1	–	–	–	–	–	3	–	–	32.50	225.00	15.00
125	–	–	1	–	–	–	–	–	2	40.00	300.00	17.32
126	–	–	–	–	–	–	5	–	1	41.67	16.67	4.08
127	1	–	–	1	–	1	1	1	–	31.00	192.50	13.87
128	–	–	–	–	3	1	1	–	–	33.00	20.00	4.47
129	–	–	–	–	–	1	1	–	1	41.67	58.33	7.64
130	–	–	–	–	–	–	1	–	2	46.67	33.33	5.77
131	–	–	–	–	–	–	1	2	1	45.00	16.67	4.08
132	–	–	–	–	2	–	3	–	1	38.33	56.67	7.53
133	–	1	1	–	2	–	2	–	–	29.17	104.17	10.21
134	–	–	1	–	1	1	–	–	–	28.33	58.33	7.64
135	–	–	–	–	1	–	3	–	1	40.00	50.00	7.07
136	–	–	–	1	1	–	1	–	–	31.67	58.33	7.64
137	–	–	–	1	4	–	–	–	–	29.00	5.00	2.24
138	–	–	3	–	–	–	–	1	–	26.25	156.25	12.50
139	1	–	–	–	2	–	–	–	–	23.33	133.33	11.55
140	–	–	1	–	1	1	–	–	–	28.33	58.33	7.64
141	–	–	–	–	2	–	–	–	–	30.00	0.00	0.00
142	1	–	–	1	2	–	1	–	1	30.83	184.17	13.57
143	–	–	–	–	–	–	3	–	–	40.00	0.00	0.00
144	–	–	–	–	1	–	2	1	1	41.00	55.00	7.42
145	1	–	–	–	1	–	1	–	–	26.67	233.33	15.28
146	–	–	–	–	1	1	1	–	–	35.00	25.00	5.00
147	–	–	1	–	–	–	3	–	–	35.00	100.00	10.00
148	–	–	–	–	1	–	2	–	–	36.67	33.33	5.77
149	–	–	2	–	–	–	1	–	–	26.67	133.33	11.55
150	1	–	2	–	2	1	–	–	–	24.17	84.17	9.17

TABLE A-4 Raw Data of Peer Reviews and Basic Statistics for the 50 Proposals in Economics: NSF Reviews

Proposal Number	Peer Review Score									Mean	Variance	Standard Deviation
	10	15	20	25	30	35	40	45	50			
301	–	–	1	–	1	1	1	–	1	35.00	125.00	11.18
302	–	–	–	–	2	–	–	–	–	30.00	0.00	0.00
303	–	–	1	1	–	–	1	–	–	28.33	108.33	10.41
304	–	–	–	–	1	–	1	–	–	35.00	50.00	7.07
305	–	–	–	–	–	1	1	–	2	43.75	56.25	7.50
306	–	–	–	–	1	–	1	–	1	40.00	100.00	10.00
307	–	–	–	–	–	–	–	–	5	50.00	0.00	0.00
308	–	–	1	–	–	–	2	–	1	37.50	158.33	12.58
309	–	–	–	–	1	–	2	–	1	40.00	66.67	8.17
310	–	–	–	–	–	–	2	–	2	45.00	33.33	5.77
311	–	–	–	–	–	–	4	–	2	43.33	26.67	5.16
312	–	–	–	–	–	–	2	–	1	43.33	33.33	5.77
313	–	–	–	–	1	–	4	–	1	40.00	40.00	6.33
314	–	–	–	–	1	–	1	–	1	40.00	100.00	10.00
315	–	–	–	–	–	–	1	–	–	40.00	0.00	0.00
316	–	–	1	–	–	–	–	–	1	35.00	450.00	21.21
317	–	–	–	1	–	–	1	–	1	38.33	158.33	12.58
318	–	–	–	–	–	–	1	1	1	45.00	25.00	5.00
319	–	–	–	–	–	–	2	–	1	43.33	33.33	5.77
320	–	–	–	–	–	–	1	–	1	45.00	50.00	7.07
321	–	–	1	–	1	–	–	–	1	33.33	233.33	15.28
322	–	–	1	–	4	–	–	–	1	31.67	96.67	9.83
323	1	–	–	–	1	–	2	–	–	30.00	200.00	14.14
324	–	–	–	–	–	–	–	–	2	50.00	0.00	0.00
325	–	–	–	–	–	–	2	–	1	43.33	33.33	5.77
326	1	–	–	–	–	–	–	–	–	10.00	0.00	0.00
327	–	–	1	–	–	–	–	–	–	20.00	0.00	0.00
328	–	–	1	–	2	–	1	–	–	30.00	66.67	8.17
329	–	–	1	–	–	–	2	–	1	37.50	158.33	12.58
330	–	–	–	–	–	1	1	–	3	45.00	50.00	7.07
331	–	–	1	–	1	–	1	–	–	30.00	100.00	10.00
332	1	–	–	–	–	–	3	–	1	36.00	230.00	15.17
333	–	–	1	–	–	–	–	–	–	20.00	0.00	0.00
334	–	–	1	–	–	–	–	–	–	20.00	0.00	0.00
335	–	–	1	–	1	–	3	–	–	34.00	80.00	8.94
336	–	–	1	–	–	–	–	–	1	35.00	450.00	21.21
337	1	–	3	–	1	–	–	–	–	20.00	50.00	7.07
338	–	–	–	–	1	–	3	–	1	40.00	50.00	7.07
339	–	–	2	–	–	–	–	–	–	20.00	0.00	0.00
340	–	–	–	1	1	–	1	–	–	31.67	58.33	7.64
341	–	–	1	–	2	–	–	–	–	26.67	33.33	5.77
342	–	–	1	–	–	–	–	–	–	20.00	0.00	0.00
343	1	–	1	–	1	–	2	–	–	28.00	170.00	13.04
344	–	–	–	–	1	–	–	–	2	43.33	133.33	11.55
345	–	–	1	–	–	–	–	–	–	20.00	0.00	0.00
346	–	–	1	–	–	–	1	–	1	36.67	233.33	15.28
347	–	–	1	1	1	–	–	–	–	25.00	25.00	5.00
348	–	–	1	–	–	–	–	–	–	20.00	0.00	0.00
349	–	–	–	–	2	–	1	–	–	33.33	33.33	5.77
350	5	–	2	–	–	–	–	–	–	12.86	23.81	4.88

TABLE A-5 Raw Data of Peer Reviews and Basic Statistics for the 50 Proposals in Economics: Reviews of Non-Blinded Proposals

Proposal Number	Peer Review Score									Mean	Variance	Standard Deviation
	10	15	20	25	30	35	40	45	50			
301	–	–	–	–	3	–	–	–	–	30.00	0.00	0.00
302	–	1	2	–	–	–	2	1	–	30.00	170.00	13.04
303	1	–	2	–	–	–	1	–	–	22.50	158.33	12.58
304	–	–	1	–	1	–	–	–	–	25.00	50.00	7.07
305	–	–	1	–	1	–	2	–	–	32.50	91.67	9.57
306	–	–	1	–	1	1	2	–	–	33.00	70.00	8.37
307	–	–	–	–	1	–	–	–	4	46.00	80.00	8.94
308	–	–	1	–	1	1	–	–	–	28.33	58.33	7.64
309	–	–	1	–	–	–	–	–	2	40.00	300.00	17.32
310	–	–	–	–	1	–	2	–	–	36.67	33.33	5.77
311	–	–	–	–	2	–	2	–	–	35.00	33.33	5.77
312	–	–	1	–	3	–	–	–	–	27.50	25.00	5.00
313	–	–	1	–	–	1	1	–	–	31.67	108.33	10.41
314	–	–	–	–	–	1	3	1	–	40.00	12.50	3.54
315	–	–	–	1	–	–	1	–	2	41.25	139.58	11.82
316	–	–	–	–	–	–	1	2	2	46.00	17.50	4.18
317	–	–	3	–	–	–	1	–	–	25.00	100.00	10.00
318	–	–	–	–	1	–	–	–	1	40.00	200.00	14.14
319	–	–	–	–	–	–	3	–	1	42.50	25.00	5.00
320	–	–	–	–	–	–	1	–	3	47.50	25.00	5.00
321	–	–	1	–	1	–	–	–	1	33.33	233.33	15.28
322	–	1	3	–	–	–	1	–	–	23.00	95.00	9.75
323	–	–	3	–	1	–	–	–	–	22.50	25.00	5.00
324	–	–	–	–	–	–	–	–	3	50.00	0.00	0.00
325	–	–	–	–	1	–	2	–	–	36.67	33.33	5.77
326	1	–	2	–	–	–	–	–	–	16.67	33.33	5.77
327	1	–	1	–	–	–	–	–	–	15.00	50.00	7.07
328	–	1	2	1	–	–	–	–	–	20.00	16.67	4.08
329	1	–	–	–	2	–	2	–	–	30.00	150.00	12.25
330	–	–	1	–	–	–	–	–	4	44.00	180.00	13.42
331	1	–	1	–	1	–	2	–	1	31.67	216.67	14.72
332	–	–	1	–	1	–	2	–	–	32.50	91.67	9.57
333	1	–	1	–	–	–	1	1	–	28.75	272.92	16.52
334	–	–	1	–	–	1	–	–	1	35.00	225.00	15.00
335	–	–	–	–	1	–	1	–	–	35.00	50.00	7.07
336	–	–	–	1	3	–	–	–	–	28.75	6.25	2.50
337	–	–	1	–	2	–	–	–	–	26.67	33.33	5.77
338	–	–	1	–	–	–	1	–	–	30.00	200.00	14.14
339	1	–	3	–	–	–	–	–	2	28.33	296.67	17.22
340	–	–	1	–	1	–	1	–	–	30.00	100.00	10.00
341	1	1	1	–	2	–	–	–	–	21.00	80.00	8.94
342	–	–	–	1	–	–	2	–	–	35.00	75.00	8.66
343	–	–	1	–	2	1	–	–	–	28.75	39.58	6.29
344	–	–	–	1	–	–	1	–	–	32.50	112.50	10.61
345	1	–	–	–	–	–	–	–	–	10.00	0.00	0.00
346	–	–	–	–	1	–	1	–	–	35.00	50.00	7.07
347	1	–	2	1	–	–	–	–	–	18.75	39.58	6.29
348	–	–	–	–	–	–	2	–	–	40.00	0.00	0.00
349	–	–	–	–	–	–	2	–	2	45.00	33.33	5.77
350	4	–	–	–	–	–	–	–	–	10.00	0.00	0.00

TABLE A-6 Raw Data of Peer Reviews and Basic Statistics for the 50 Proposals in Economics: Reviews of Blinded Proposals

Proposal Number	Peer Review Score									Mean	Variance	Standard Deviation
	10	15	20	25	30	35	40	45	50			
301	–	–	3	–	1	–	–	–	–	22.50	25.00	5.00
302	–	–	3	–	1	–	–	–	1	28.00	170.00	13.04
303	1	–	1	–	3	–	–	–	–	24.00	80.00	8.94
304	–	–	2	–	–	–	–	–	2	35.00	300.00	17.32
305	–	–	2	–	2	–	–	–	–	25.00	33.33	5.77
306	–	–	1	–	–	–	1	–	–	30.00	200.00	14.14
307	–	–	–	–	–	–	2	–	3	46.00	30.00	5.48
308	–	1	–	–	2	–	1	–	1	33.00	170.00	13.04
309	–	–	1	–	1	–	1	–	2	38.00	170.00	13.04
310	–	–	–	–	1	–	1	–	3	44.00	80.00	8.94
311	–	–	–	–	–	–	2	–	–	40.00	0.00	0.00
312	–	–	1	–	1	–	–	–	2	37.50	225.00	15.00
313	–	–	1	–	–	–	–	–	–	20.00	0.00	0.00
314	–	–	4	1	1	–	–	–	–	22.50	17.50	4.18
315	–	–	–	–	3	–	2	–	1	36.67	66.67	8.17
316	–	–	–	–	–	–	–	–	5	50.00	0.00	0.00
317	–	–	2	–	–	–	1	–	1	32.50	225.00	15.00
318	–	–	–	–	1	–	2	–	–	36.67	33.33	5.77
319	–	–	3	–	–	–	1	–	–	25.00	100.00	10.00
320	–	–	–	–	2	–	1	–	2	40.00	100.00	10.00
321	–	–	1	–	1	–	1	–	–	30.00	100.00	10.00
322	1	–	1	–	–	–	1	–	–	23.33	233.33	15.28
323	1	–	3	–	–	–	–	–	–	17.50	25.00	5.00
324	–	–	–	–	–	–	1	1	2	46.25	22.92	4.79
325	–	–	2	–	–	–	1	–	–	26.67	133.33	11.55
326	3	–	–	–	–	–	–	–	–	10.00	0.00	0.00
327	1	–	3	–	–	–	–	–	–	17.50	25.00	5.00
328	–	–	–	–	2	–	3	–	–	36.00	30.00	5.48
329	–	–	–	–	3	–	1	–	1	36.00	80.00	8.94
330	–	–	1	–	2	–	–	–	–	26.67	33.33	5.77
331	2	–	–	–	–	–	2	–	–	25.00	300.00	17.32
332	1	–	–	1	2	–	–	–	–	23.75	89.58	9.47
333	1	–	3	–	2	–	–	–	–	21.67	56.67	7.53
334	–	–	2	–	–	–	2	–	–	30.00	133.33	11.55
335	–	–	1	–	2	1	1	–	–	31.00	55.00	7.42
336	1	–	3	–	–	–	–	–	–	17.50	25.00	5.00
337	1	1	1	–	1	–	–	–	–	18.75	72.92	8.54
338	–	–	1	1	1	–	1	–	–	28.75	72.92	8.54
339	–	–	1	–	–	–	3	–	1	38.00	120.00	10.95
340	–	–	2	–	–	–	1	–	–	26.67	133.33	11.55
341	–	2	–	–	1	–	1	–	1	30.00	237.50	15.41
342	–	–	1	–	1	–	–	–	–	25.00	50.00	7.07
343	–	–	2	–	–	–	1	–	1	32.50	225.00	15.00
344	2	–	2	–	–	–	1	–	–	20.00	150.00	12.25
345	1	–	3	–	1	–	–	–	–	20.00	50.00	7.07
346	–	–	1	1	–	–	1	–	–	28.33	108.33	10.41
347	1	1	–	–	–	–	–	–	–	12.50	12.50	3.54
348	–	–	1	–	1	–	–	–	1	33.33	233.33	15.28
349	–	–	1	–	1	–	1	–	–	30.00	100.00	10.00
350	3	–	–	–	–	–	–	–	–	10.00	0.00	0.00

TABLE A-7 Raw Data of Peer Reviews and Basic Statistics for the 50 Proposals in Solid-State Physics: NSF Reviews

Proposal Number	Peer Review Score									Mean	Variance	Standard Deviation
	10	15	20	25	30	35	40	45	50			
201	–	–	–	–	–	–	1	–	1	45.00	50.00	7.07
202	–	–	3	–	–	1	–	–	–	23.75	56.25	7.50
203	–	–	–	–	–	–	–	–	2	50.00	0.00	0.00
204	–	–	–	–	–	–	1	–	3	47.50	25.00	5.00
205	–	–	–	–	–	–	2	–	1	43.33	33.33	5.77
206	–	–	–	–	–	–	1	1	–	42.50	12.50	3.54
207	1	–	–	–	–	1	2	–	–	31.25	206.25	14.36
208	–	–	–	–	2	–	2	–	–	35.00	33.33	5.77
209	–	–	–	–	–	1	2	2	–	41.00	17.50	4.18
210	–	–	–	–	–	1	–	1	–	40.00	50.00	7.07
211	–	–	–	–	–	1	2	–	1	41.25	39.58	6.29
212	–	–	–	–	1	–	3	–	–	37.50	25.00	5.00
213	–	–	–	–	–	–	3	–	2	44.00	30.00	5.48
214	–	–	–	–	1	2	1	–	–	35.00	16.67	4.08
215	–	–	–	–	1	–	3	1	–	39.00	30.00	5.48
216	–	–	1	1	1	–	–	1	–	30.00	116.67	10.80
217	–	–	–	–	1	–	3	–	–	37.50	25.00	5.00
218	–	–	–	–	–	–	2	1	1	43.75	22.92	4.79
219	–	–	–	–	–	–	2	1	–	41.67	8.33	2.89
220	–	–	–	–	–	1	2	–	2	43.00	45.00	6.71
221	–	–	–	–	1	–	3	1	–	39.00	30.00	5.48
222	–	–	2	–	–	–	3	–	–	32.00	120.00	10.95
223	–	–	–	–	1	–	3	–	–	37.50	25.00	5.00
224	–	–	–	–	1	–	2	2	2	42.86	48.81	6.99
225	–	–	–	–	–	–	4	–	–	40.00	0.00	0.00
226	–	–	–	–	1	–	1	1	1	41.25	72.92	8.54
227	–	–	–	–	–	–	1	–	2	46.67	33.33	5.77
228	–	–	1	–	1	1	–	–	–	28.33	58.33	7.64
229	–	–	–	–	1	–	2	–	1	40.00	66.67	8.17
230	–	–	–	–	–	–	2	1	1	43.75	22.92	4.79
231	–	–	–	–	–	2	1	–	–	36.67	8.33	2.89
232	–	–	1	–	–	1	2	–	1	37.00	120.00	10.95
233	–	–	3	–	1	–	–	–	–	22.50	25.00	5.00
234	–	–	1	–	1	–	1	–	1	35.00	166.67	12.91
235	–	–	–	–	1	1	1	–	–	35.00	25.00	5.00
236	–	–	–	–	–	–	3	–	–	40.00	0.00	0.00
237	–	–	–	–	–	–	2	–	1	43.33	33.33	5.77
238	–	–	1	–	2	–	1	–	–	30.00	66.67	8.17
239	–	–	–	–	–	–	3	–	1	42.50	25.00	5.00
240	–	–	–	–	1	1	2	–	–	36.25	22.92	4.79
241	–	–	1	–	1	1	–	–	–	28.33	58.33	7.64
242	–	–	–	–	–	1	3	–	–	38.75	6.25	2.50
243	–	–	1	–	1	–	2	–	–	32.50	91.67	9.57
244	–	–	–	–	1	1	2	–	–	36.25	22.92	4.79
245	–	–	–	–	1	–	2	–	1	40.00	66.67	8.17
246	–	–	–	–	–	–	3	–	2	44.00	30.00	5.48
247	–	–	1	–	2	–	1	–	–	30.00	66.67	8.17
248	–	–	–	–	–	1	1	–	1	41.67	58.33	7.64
249	–	–	–	–	–	–	1	–	3	47.50	25.00	5.00
250	–	–	1	–	1	–	1	–	–	30.00	100.00	10.00

TABLE A-8 Raw Data of Peer Reviews and Basic Statistics for the 50 Proposals in Solid-State Physics: Reviews of Non-Blinded Proposals

Proposal Number	Peer Review Score									Mean	Variance	Standard Deviation
	10	15	20	25	30	35	40	45	50			
201	–	–	–	–	–	–	–	–	4	50.00	0.00	0.00
202	1	–	2	–	–	–	–	–	–	16.67	33.33	5.77
203	–	–	–	–	–	–	1	–	3	47.50	25.00	5.00
204	–	–	–	–	1	–	2	–	1	40.00	66.67	8.17
205	–	–	–	–	1	–	–	–	–	30.00	0.00	0.00
206	–	–	–	–	1	1	2	1	–	38.00	32.50	5.70
207	–	–	1	–	2	–	1	–	–	30.00	66.67	8.17
208	–	–	1	–	4	–	–	–	–	28.00	20.00	4.47
209	–	–	–	–	1	1	2	–	–	36.25	22.92	4.79
210	–	–	–	–	–	–	1	1	2	46.25	22.92	4.79
211	–	–	–	–	2	–	1	–	1	37.50	91.67	9.57
212	–	–	–	–	2	–	–	1	–	35.00	75.00	8.66
213	–	–	–	–	–	–	3	–	3	45.00	30.00	5.48
214	–	–	1	–	1	–	2	–	–	32.50	91.67	9.57
215	–	–	–	–	–	1	1	–	2	43.75	56.25	7.50
216	–	–	1	–	–	1	1	–	–	31.67	108.33	10.41
217	–	–	–	–	4	–	–	–	–	30.00	0.00	0.00
218	–	–	–	–	1	1	1	–	–	35.00	25.00	5.00
219	–	–	–	–	–	–	2	–	1	43.33	33.33	5.77
220	–	–	–	–	–	–	4	–	–	40.00	0.00	0.00
221	–	–	–	–	–	–	1	–	1	45.00	50.00	7.07
222	–	–	–	–	1	–	1	–	–	35.00	50.00	7.07
223	–	1	–	–	1	–	3	1	–	35.00	120.00	10.95
224	–	–	–	–	–	–	2	2	–	42.50	8.33	2.89
225	–	–	–	–	1	–	1	–	–	35.00	50.00	7.07
226	–	–	–	–	2	–	2	–	2	40.00	80.00	8.94
227	–	–	1	–	2	–	1	–	1	34.00	130.00	11.40
228	–	–	1	–	–	–	1	–	–	30.00	200.00	14.14
229	–	–	–	–	1	–	3	–	1	40.00	50.00	7.07
230	–	–	–	–	2	1	1	–	1	37.00	70.00	8.37
231	–	–	1	–	–	–	1	–	–	30.00	200.00	14.14
232	–	–	1	–	2	–	1	–	–	30.00	66.67	8.17
233	1	–	–	–	2	1	–	–	–	26.25	122.92	11.09
234	–	–	–	–	1	1	1	–	–	35.00	25.00	5.00
235	–	–	1	1	2	–	–	–	–	26.25	22.92	4.79
236	–	–	–	–	2	1	1	–	–	33.75	22.92	4.79
237	–	–	–	–	2	–	2	–	2	40.00	80.00	8.94
238	–	–	1	–	4	–	–	–	–	28.00	20.00	4.47
239	–	–	–	–	–	2	2	–	–	37.50	8.33	2.89
240	–	–	–	–	2	–	1	–	–	33.33	33.33	5.77
241	–	–	1	–	–	–	1	–	–	30.00	200.00	14.14
242	–	–	–	–	3	–	–	–	–	30.00	0.00	0.00
243	–	–	–	–	1	–	3	–	1	40.00	50.00	7.07
244	–	–	–	–	–	3	2	–	–	37.00	7.50	2.74
245	–	–	–	–	2	–	2	–	–	35.00	33.33	5.77
246	–	–	–	–	1	1	–	–	–	32.50	12.50	3.54
247	–	–	–	–	1	–	1	1	–	38.33	58.33	7.64
248	–	–	–	–	–	–	3	–	–	40.00	0.00	0.00
249	–	–	1	–	1	2	1	–	–	32.00	57.50	7.58
250	–	–	–	–	4	–	–	–	–	30.00	0.00	0.00

TABLE A-9 Raw Data of Peer Reviews and Basic Statistics for the 50 Proposals in Solid-State Physics: Reviews of the Blinded Proposals

Proposal Number	Peer Review Score									Mean	Variance	Standard Deviation
	10	15	20	25	30	35	40	45	50			
201	–	–	–	–	–	–	–	–	5	50.00	0.00	0.00
202	1	–	1	–	1	–	–	–	–	20.00	100.00	10.00
203	–	–	–	–	–	–	2	–	1	43.33	33.33	5.77
204	–	–	–	–	1	1	2	–	–	36.25	22.97	4.79
205	–	–	–	–	–	1	1	–	1	41.67	58.33	7.64
206	–	–	2	–	–	–	1	–	–	26.67	133.33	11.55
207	1	–	–	–	1	–	–	–	1	30.00	400.00	20.00
208	–	–	–	–	3	–	3	–	–	35.00	30.00	5.48
209	–	–	1	–	1	–	1	1	–	33.75	122.92	11.09
210	–	–	–	–	2	–	3	–	–	36.00	30.00	5.48
211	–	–	–	–	–	–	3	–	1	42.50	25.00	5.00
212	–	–	2	–	–	2	–	–	–	27.50	75.00	8.66
213	–	–	–	–	–	–	2	1	3	45.83	24.17	4.92
214	–	–	–	–	4	–	1	–	–	32.00	20.00	4.47
215	–	–	–	–	1	1	1	–	2	41.00	80.00	8.94
216	–	–	–	1	–	–	2	–	–	35.00	75.00	8.66
217	–	–	1	–	1	–	–	–	–	25.00	50.00	7.07
218	–	–	–	–	2	–	1	–	1	37.50	91.67	9.57
219	–	–	–	–	–	–	2	1	–	41.67	8.33	2.89
220	–	–	–	–	1	–	2	–	–	36.67	33.33	5.77
221	–	–	–	–	1	–	–	–	2	43.33	133.33	11.55
222	–	–	1	–	–	–	2	–	1	37.50	158.33	12.58
223	1	–	2	–	1	–	1	–	–	24.00	130.00	11.40
224	–	–	–	–	–	–	2	1	1	43.75	22.92	4.79
225	–	1	1	–	–	–	1	–	–	25.00	175.00	13.23
226	–	–	1	–	2	–	–	–	–	26.67	33.33	5.77
227	–	–	–	–	1	1	1	–	3	42.50	77.50	8.80
228	1	–	2	–	1	–	–	–	–	20.00	66.67	8.17
229	1	–	–	–	2	–	2	–	1	33.33	186.67	13.66
230	–	–	–	–	–	–	1	–	2	46.67	33.33	5.77
231	–	–	1	–	2	–	1	–	–	30.00	66.67	8.17
232	–	–	–	–	1	1	–	–	3	43.00	95.00	9.75
233	–	–	1	–	3	–	1	–	–	30.00	50.00	7.07
234	–	–	–	–	4	–	–	–	–	30.00	0.00	0.00
235	–	–	1	–	2	–	1	–	–	30.00	66.67	8.17
236	1	–	1	1	–	–	1	–	–	23.75	156.25	12.50
237	–	–	–	–	2	–	1	–	1	37.50	91.67	9.57
238	–	–	3	–	–	–	1	–	–	25.00	100.00	10.00
239	–	–	–	–	1	1	2	–	–	36.25	22.91	4.79
240	–	–	–	1	1	–	–	–	1	35.00	175.00	13.23
241	–	–	–	1	–	–	2	–	1	38.75	106.25	10.31
242	–	–	1	–	2	–	–	–	–	26.67	33.33	5.77
243	–	–	1	–	2	–	1	–	1	34.00	130.00	11.40
244	–	–	–	–	4	–	1	–	–	32.00	20.00	4.47
245	–	–	–	–	1	–	2	–	1	40.00	66.67	8.17
246	–	–	–	–	–	–	2	–	2	45.00	33.33	5.77
247	–	–	–	–	2	–	3	1	–	37.50	37.50	6.12
248	–	–	–	–	–	–	3	–	1	42.50	25.00	5.00
249	–	–	–	–	–	–	2	–	2	45.00	33.33	5.77
250	–	–	–	–	1	–	3	–	–	37.50	25.00	5.00

APPENDIX B

DOCUMENTS USED IN THE COLLECTION OF THE PHASE II DATA SET

I am writing to request your aid in a research project being conducted by the Committee on Science and Public Policy (COSPUP), National Academy of Sciences.

Congress recently mandated that the National Science Foundation conduct research on whether peer review of proposals submitted to NSF could be conducted without revealing the identity of the Principal Investigator. The NSF requested that COSPUP include this task in its general study of the peer review system at the Foundation which has been underway for the past 2 years.

To comply with the Congressional mandate, we are conducting an experiment on several NSF programs. We have selected a random sample of proposals on which the NSF has recently made a decision. Then, using a panel of experts in the field, we selected a group of new reviewers. The chairman of the review selector committee for . . . *_____ The panel was asked to identify individuals who would be particularly well suited to review specific proposals.

---

*The blank space above allows for one of the three following inserts:

(1) . . . solid state physics was Dr. Albert M. Clogston of Bell Laboratories.

(2) . . . solid state physics was Dr. George Pimentel of Berkeley.

(3) . . . economics was Dr. Harold W. Watts of Columbia University.

You were one of the reviewers selected. Half of the reviewers are being asked to review identified proposals and the other half, selected at random, are being asked to review the same proposals unidentified. We will then compare the ratings received from our two sets of reviewers with those obtained by the NSF reviewers.

We would greatly appreciate it if you could review the enclosed proposal and give it a rating from excellent to poor. It would also help us if you could briefly give the reasons for your rating.

In many cases it is difficult or impossible to conceal the identity of the Principal Investigator. If you think you know who the Principal Investigator is, please indicate this on the appropriate line on the review sheet. In any case, please try to base your evaluation of the proposal strictly on the quality of the science proposed.

We would appreciate it if you could return your review in the enclosed stamped envelope. Please destroy the proposal. If for any reason you will be unable to complete your evaluation within the next month, please return the proposal to the Academy.

Your review of this proposal will be treated confidentially. the specific reviews received will not be conveyed to the NSF and will have no influence on the actual granting decision which has already been made. Your cooperation will, however, enable us to complete what we believe is an important research project.

Thanking you in advance for your cooperation.

Sincerely,

I. M. Singer
Chairman, COSPUP

IMS:gl

Encs.

NATIONAL ACADEMY OF SCIENCE PEER REVIEW STUDY

PROPOSAL RATING SHEET

Reviewer:_____ Proposal No.:_____

Rating:

Excellent Very Good Good Fair Poor

Comments: (Continue on additional sheet if necessary)

Name of Principal Investigator(s), if you know who it is:_____

Has the removal of references from this proposal made it difficult to understand?

Yes No

Is it more or less difficult to review a proposal on which an attempt has been made to conceal the identity of the Principal Investigator?

more difficult | less difficult | does not influence the difficulty

In deciding who should get NSF grants, do you think that more weight should be given to the content of the proposal or the past research performance of the Principal Investigator?

Content of Proposal
Past research performance of Principal Investigator
Both about the same

I am writing to request your aid in a research project being conducted by the Committee on Science and Public Policy, National Academy of Sciences. For the past two years we have been studying the peer review system employed by the National Science Foundation. As part of our research we are conducting an experiment on several NSF programs.

Within these programs we have selected a random sample of proposals on which the NSF has recently made a decision. Then, using a panel of experts in the field, we selected a group of new reviewers. The chairman of the review selector committee for . . .*_____ The panel was asked to identify individuals who would be particularly well suited to review specific proposals. You were one of the reviewers selected. The primary aim of the experiment is to compare the reviews obtained from our independently selected reviewers with those obtained from the NSF reviewers. We would greatly appreciate it if you could review the enclosed proposal and give it a rating from excellent to poor. It would also help us if you could briefly give the reasons for your rating.

The following definitions of the overall rating terms are supplied for your guidance in evaluating the attached proposal.

Excellent: Presents an opportunity for a major contribution to solid state physics; is in the upper ten percent of research proposals; should be given the highest priority for funding.

Very Good: A superior proposal; will make an important contribution to solid state physics and should be supported.

Good: May make a contribution to solid state physics; scientifically acceptable but

*The blank space above allows for one of the three following inserts:

(1) . . . solid state physics was Dr. Albert M. Clogston of Bell Laboratories.

(2) . . . solid state physics was Dr. George Pimentel of Berkeley.

(3) . . . economics was Dr. Harold W. Watts of Columbia University.

may border on the routine or unimaginative; low priority for funding.

Fair: Probability for an important contribution to solid state physics is low; not deserving of support.

Poor: Unsatisfactory, poorly planned or of a purely routine character; not deserving of support.

Your review should reflect primarily the merit of the proposed research including detailed comments on the technical aspects of the proposal, the competence of the investigator, and the relevant resources of the institution. Comments on the significance of the proposed research for its immediate field, and more generally for science, as well as on the extent to which the proposed work duplicates or overlaps other existing work, would be helpful. Proposals for renewal of an existing grant should be reviewed also in terms of previous progress and productivity.

Reviewers may be chosen from several fields of science to help us judge the significance and promise of the proposed work. We need your impressions of the proposal as a whole as well as your detailed comments.

We would appreciate it if you could return your review in the enclosed stamped envelope. Please destroy the proposal. If for any reason you will be unable to complete your evaluation within the next month, please return the proposal to the Academy.

Your review of this proposal will be treated confidentially. The specific reviews received will not be conveyed to the NSF and will have no influence on the granting decision which has already been made. Your cooperation will, however, enable us to complete what we believe is an important research project.

Thanking you in advance for your cooperation.

Sincerely,

IMS:gl

I. M. Singer
Chairman, COSPUP

Encs.

NATIONAL ACADEMY OF SCIENCES PEER REVIEW STUDY

PROPOSAL RATING SHEET

Reviewer:_____ Proposal No.:_____

Investigator:_____ Institution: _____

Comments: (Continue on additional sheet if necessary)

OVERALL RATING:

Excellent

Very Good

Good

Fair

Poor

September, 1978

Several weeks ago we wrote to you and asked for your continuing cooperation in a study currently being conducted by the Committee on Science and Public Policy of the National Academy of Sciences. We requested that you take a minute or two in answering two questions that ask for appraisals of the contributions that the principal investigator has made to his or her field over the past ten years. To date, we believe we have not yet received a reply from you on this latest inquiry. If you have already completed this questionnaire, please disregard this reminder.

It is only through the cooperation of scientists like yourself that we shall be able to complete what we believe to be an important study. It would be of great assistance to us and very much appreciated if you could take the time to answer these two questions. Enclosed please find original correspondence, questionnaire, and self-addressed stamped envelope for your convenience.

Once again, thank you for your cooperation.

Sincerely yours,

Stephen Cole
Co-Project Director

Encs.

Jonathan R. Cole
Co-Project Director

Committee on Science and Public Policy
National Academy of Sciences
2101 Constitution Avenue
Washington, D.C. 20418

A month or two ago I wrote you and asked for your cooperation in a study currently being conducted by the Committee on Science and Public Policy of the National Academy of Sciences. We are studying how the peer review system currently employed by the National Science Foundation operates. As part of this study we are conducting an experiment to see whether or not an independently selected group of reviewers would evaluate proposals in the same way as reviewers selected by the NSF. Your name was selected as a reviewer of one or two proposals by our Reviewer Selector Committee. To date, we have not yet received a reply from you.

Since there is a relatively small number of qualified reviewers for each proposal, we would greatly appreciate it if you could review the proposal(s) we sent you and return the reviews to us no later than December 15, 1977. (If the proposal(s) we sent you has been misplaced, please write or phone us and we will send you new copies.) If it will be impossible for you to review the proposal(s) we sent you, we would greatly appreciate it if you could return them to the National Academy of Sciences.

I would like to thank you in advance for your cooperation in what we believe to be an important study.

Sincerely yours,

I. M. Singer
Chairman, COSPUP

Committee on Science and Public Policy
National Academy of Sciences
2101 Constitution Avenue
Washington, D.C. 20418

A month or two ago I wrote you and asked for your cooperation in a study that the National Academy of Sciences is currently conducting of the National Science Foundation peer review system. As part of this study we have, for a selected group of proposals, chosen new reviewers and have asked them to review one or two proposals in the general area of their expertise. At the time you received the proposal you were unable to review it for us immediately. Since there is a relatively small number of qualified reviewers for each proposal we are, once again, asking for your help in completing what we believe to be an important research project.

If it will be impossible for you to review the enclosed proposal(s) prior to December 15, 1977, then we would greatly appreciate it if you could return it to the National Academy of Sciences.

Thanking you in advance for your cooperation.

Sincerely yours,

I. M. Singer
Chairman, COSPUP

Committee on Science and Public Policy
National Academy of Sciences
2101 Constitution Avenue
Washington, D.C. 20418

In the last several months we have written to you and asked for your cooperation in a study currently being conducted by the Committee on Science and Public Policy of the National Academy of Sciences. We are studying how the peer review system employed by the National Science Foundation operates. As part of this study we are conducting an experiment to see whether or not an independently selected group of reviewers would evaluate proposals in the same way as reviewers selected by the NSF. Your name was selected as a reviewer of one or two proposals by our Reviewer Selector Committee. To date, we have not yet received a reply from you.

We are having some difficulty in completing this study since there is only a relatively small number of qualified reviewers for each proposal. It is only through the cooperation of scientists like yourself that we shall be able to complete what we believe to be an important study. It would be of great assistance to us and very much appreciated if you could find the time to review the proposal(s) which we sent you and return your reviews to us no later than February 15, 1978. We will be unable to use reviews which are not returned by this date, since we must complete our analysis shortly thereafter.

If the proposal(s) we sent you has been misplaced, please write or phone us collect at the office of the Project Director, Dr. Stephen Cole, 516-246-8681. If it will be impossible for you to review the proposal(s) we sent you, we would appreciate it if you could return it to the National Academy of Sciences.

I would like to thank you in advance for your cooperation.

Sincerely yours,

I. M. Singer
Chairman, COSPUP

Committee on Science and Public Policy
National Academy of Sciences
2101 Constitution Avenue
Washington, D.C. 20418

July 22, 1977

At the suggestion of Dr. Albert M. Clogston of Bell Laboratories I am once again writing to you to ask for your help in the study we are conducting of the National Science Foundation peer review system. Thanks to your cooperation and that of other reviewer selectors, we now have lists of potential reviewers for each of the 50 proposals we are studying in solid state physics. However, due to several problems, we do not have quite enough reviewers for some of the proposals. These problems were as follows:

1) Since the proposals did not identify the principal investigator, or his institution, it was not infrequent that a reviewer selector would select as a reviewer the principal investigator or even more frequently other scientists located at the same institution as the principal investigator. We cannot, of course, use these people as reviewers.

2) Since we had two reviewer selectors for each proposal it was not unusual for both reviewer selectors to mention several names in common. This further reduced the number of useable reviewers.

3) It was also not unusual for one or the other of the reviewer selectors to choose as a reviewer for a particular proposal a reviewer whom the National Science Foundation has already employed. Since we want our review of the proposals to be completely independent of that conducted by the Foundation, we can not use these reviewers.

Enclosed with this letter are sheets which contain the name of the proposal and indicate the institution of the principal investigator. We also list all the reviewers we currently have for this proposal, including those employed by the Foundation. For each proposal, we indicate the number of additional new reviewers that we need in order to complete the experiment. Generally, we need only two or three additional names. In order to help you select these additional reviewers, we are enclosing a list of references taken from the proposal.

Both Dr. Clogston and I know how busy you are, and we hesitate to ask you for this additional aid. However,

without the help of knowledgeable scientists like yourself, we could not complete what we believe to be an important study in the area of science policy.

Once again, we thank you for your cooperation.

Sincerely yours,

Stephen Cole
Professor of Sociology

SC:mw
Enc.

Department of Sociology
State University of New York
at Stony Brook
Stony Brook, New York 11794

TITLE OF
PROPOSAL _____

_____

P.I.'S
INSTITUTION _____

NUMBER OF REVIEWERS NEEDED _____

NSF AND COSPUP REVIEWERS:

_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____

ADDITIONAL REVIEWERS:

NAME: INSTITUTION

_____

_____

_____

_____

_____

July, 1978

Several months ago we asked for your help in conducting a study of the peer review system at the National Science Foundation. We sent you a scientific proposal, which had been submitted to the NSF and had recently been evaluated by their methods of peer review. We asked for your evaluation of the proposal as part of an independent test of the adequacy of the NSF methods. Roughly 75% of the scientists who were asked to participate in the experiment returned evaluations of the proposals.

Work on this study is nearing completion and the results will be reported in a National Academy monograph. Without your help the study simply could not have been done. We thank you for your cooperation in a study that we believe will produce important policy recommendations.

There remains one feature of the study that we hope you can help us with. On the page attached to this letter we have listed the name of the principal investigator and the title of the proposal that you were sent earlier. We would appreciate it if you would spend a minute or two in answering the two questions on the attachment that asks for appraisals of the contributions that the principal investigator has made to his or her field over the past ten years. Your responses will remain strictly confidential and responses will be used only in the form of statistical summaries or in a way that will not identify either you or the scientist whose contribution is being appraised. A self-addressed return envelope is enclosed for you to return the brief questionnaire.

Thank you again for your willingness to participate in this research effort.

Sincerely yours,

Stephen Cole
Co-Project Director

SC/JRC:gl

Encs.

Jonathan R. Cole
Co-Project Director

Committee on Science and Public Policy
National Academy of Sciences
2101 Constitution Avenue
Washington, D.C. 20418

National Academy of Sciences Peer Review Study

We have listed below the name of the principal investigator and the title of the proposal that you evaluated. We now ask you to give us your evaluation of the principal investigator's contribution to his scientific field and specialty through his research publications over the past 10 years. There are five categories to summarize your appraisal plus two categories that indicate that you either do not know the principal investigator's work well enough to evaluate it or that you have not heard of the principal investigator. Comments about the bases of your evaluation are welcome, but feel free to omit them. Please place an "X" next to your evaluation of the quality of the published contributions of the principal investigator.

Name of Principal Investigator: _____

_____ Proposal No._____

Title of
Proposal:_____

1. By placing an "X" next to the appropriate appraisal, please indicate your assessment of the importance over the past 10 years of the published work of the principal investigator named above relative to other members of his entire field (not limited to his specific specialty areas).

   ( ) Is among the few most important contributors to the field.
   ( ) Has made very important contributions to the field.
   ( ) Has made above average contributions to the field.
   ( ) Has made average contributions to the field.
   ( ) I am unfamiliar with the principal investigator's work but have heard of him.
   ( ) I have never heard of the principal investigator.

Comments on Evaluation: (OVER)

2. By placing an "X" next to the appropriate appraisal, please indicate your assessment of the importance over the past 10 years of the published work of the principal investigator named above relative to other members of the field working in the same research area as the evaluated proposal--that is, in the specific specialty area covered by the proposal.

( ) Is among the few most important contributors to the specialty
( ) Has made very important contributions to the specialty.
( ) Has made above average contributions to the specialty.
( ) Has made average contributions to the specialty.
( ) Has made below average contributions to the specialty.
( ) I am unfamiliar with the principal investigator's work but have heard of him or her.
( ) I have never heard of the principal investigator.

Comments on Evaluation:

Evaluator's Name (Please Print):_____

Thank you.

Please place this attachment in the enclosed return envelope.

NATIONAL SCIENCE FOUNDATION
Washington, D.C. 20550

Criteria for the Selection of Research Projects Submitted to the Division of Chemistry

Mail reviews play a key role in our evaluation of the scientific merit of proposals. The following definitions of the overall rating terms are supplied for your guidance in evaluating the attached proposal.

Excellent - Presents an opportunity for a major contribution to chemistry; is in the upper ten percent of research proposals; should be given the highest priority for funding.

Very Good - A superior proposal; will make an important contribution to chemistry; should be supported.

Good - May make a contribution to chemistry; scientifically acceptable but may border on the routine or unimaginative; low priority for funding.

Fair - Probability for an important contribution to chemistry is low; not deserving of support.

Poor - Unsatisfactory, poorly planned or of a purely routine character; not deserving of support.

Your review should reflect primarily the merit of the proposed research including detailed comments on the technical aspects of the proposal, the competence of the investigator, and the relevant resources of the institution. Comments on the significance of the proposed research for its immediate field, and more generally for science, as well as on the extent to which the proposed work duplicates or overlaps other existing work, would be helpful. Proposals for renewal of an existing grant should be reviewed also in terms of previous progress and productivity.

Reviewers may be chosen from several fields of science to help us judge the significance and promise of the proposed work. We need your impressions of the proposal as a whole as well as your detailed comments.

Note: "Criteria for the Selection of Research Projects by the National Science Foundation" as approved by the National Science Board at its 167th Meeting, October 17-18, 1974, is available on request.

Foundation Policy on Confidentiality of Proposals and Peer Reviews

The Foundation receives proposals in confidence and is responsible for protecting the confidentiality of their contents. For this reason we ask that you refrain from copying, quoting or otherwise using material from this proposal. If you believe that a colleague can make a substantive contribution to the review, please consult me before disclosing either the contents of the proposal or the applicant's name.

Verbatim copies of reviews, ratings, and associated correspondence will be sent to the principal investigator/project director on request. The copies will not contain your name, the name of your institution or names which might constitute an invasion of the privacy of others. Subject to this Foundation policy and applicable laws, including the Freedom of Information Act, 5 USC 552, your participation as a reviewer and the content of your review will be given the maximum protection from disclosure.

The Foundation will publish annually a list of the names and addresses of persons who have reviewed proposals. Individuals will not, however, be identified with specific proposals. In this way the Foundation can publicly acknowledge your service as a reviewer and at the same time protect the confidentiality of your comments.

Please destroy your copy of the proposal after your have returned your review.

B-27

NATIONAL SCIENCE FOUNDATION

PROPOSAL RATING SHEET

NSF FORM 990 TEST (1-77)

PROPOSAL NO.	INSTITUTION	PLEASE RETURN BY
PRINCIPAL INVESTIGATOR	NSF PROGRAM	
TITLE		

COMMENTS (CONTINUE ON ADDITIONAL SHEET(S) AS NECESSARY)

RATING: ☐ EXCELLENT ☐ VERY GOOD ☐ GOOD ☐ FAIR ☐ POOR

Verbatim but anonymous copies of reviews will be sent only to the principal investigator/project director on request. Subject to this NSF policy and applicable laws, including the Freedom of Information Act, 5 USC 552, reviewers' comments will be given maximum protection from disclosure.

REVIEWER'S SIGNATURE	REVIEWER'S NAME (TYPED)
OTHER SUGGESTED REVIEWERS (OPTIONAL)	

NATIONAL SCIENCE FOUNDATION
Washington, D.C. 20550

To Reviewers of Proposals:

Your review of the enclosed proposal is requested to assist us in reaching a decision concerning its support. Although you are free to use your own style of comment, we enclose a summary of the pertinent criteria for judging proposals. Your comments on the proposal are very much needed and are very carefully read. A rating sheet and a franked envelope for the return of your evaluation also are enclosed.

The Foundation receives proposals in confidence and is responsible for protecting the confidentiality of their contents. For this reason, we ask that you refrain from copying, quoting or otherwise using material from this proposal. If you believe that a colleague can make a substantive contribution to the review, please consult me before disclosing either the contents of the proposal or the applicant's name. When you have completed your review or if, for some reason, you find yourself unable to respond to this request, destroy the proposal. We also solicit your suggestions of other qualified reviewers for this proposal.

Verbatim copies of reviews, ratings and associated correspondence will be sent to the principal investigator/project director on request. The copies will not contain your name, the name of your institution or names which might constitute an invasion of the privacy of others. Subject to this Foundation policy and applicable laws, including the Freedom of Information Act, 5 USC 552, your participation as a reviewer and the content of your review will be given the maximum protection from disclosure. We therefore ask that your review be typed, if possible.

The Foundation will publish annually a list of the names and addresses of persons who have reviewed proposals. Individuals will not, however, be identified with specific proposals. In this way the Foundation can publicly acknowledge your service as a reviewer and at the same time protect the confidentiality of your comments.

If you have any questions about the proposal, please call or write. You can reach me at (202) 632-7404.

I wish to thank you personally for your help in reviewing the enclosed proposal. Your comments and ratings are a key factor in our overall assessment of research ideas presented by the scientific community. The review systems can function only if we have candid assessments of proposed research from knowledgeable scientists such as you. We very much appreciate the time and thought that go into preparing these reviews.

Sincerely yours,

D. L. Mitchell
Program Director
Solid State Physics Program
Condensed Matter Sciences Section
Division of Materials Research

NATIONAL SCIENCE FOUNDATION
Washington, D.C. 20550

Criteria for Evaluating Proposals Submitted to the Division for Materials Research

Mail reviewers are a key element in our determination of the relative scientific merit of proposals. The NSF has published detailed criteria used in evaluating proposals; excerpts are attached. As always, the most important consideration is (B) the probability that research will lead to important discoveries, generalizations, or improvements of method. A necessary prerequisite is (A) technical adequacy, while (C) utility or relevance, and (D) relationship to long term scientific potential of the U.S., may be important in certain cases.

* * *

While your rating of the proposal (see below) is a concise summary of your evaluation, we very much need and appreciate comments on the factors leading to the rating. The comments as well as the rating determine whether a proposal is funded.

Comments on the budget are appreciated because budgets often are adjusted prior to final action. However, budget considerations should not affect your rating.

Rating Terms

The rating terms mean:

Excellent: Important research that might result in a major or unique contribution to fundamental knowledge. An outstanding investigator and a well-chosen problem. Should be given highest priority for support.

Very Good: A competent investigator and a worthwhile problem. Research that can be expected to lead to substantial advances in fundamental knowledge. Should be supported.

Good: A competent investigator proposing a routine problem or a problem not well-matched to the

investigator's skill and experience. Support should be considered if funds are available.

Fair: Serious deficiencies appear to exist either in the scientific quality of the proposal or in the qualifications of the investigator. Successful completion of the research is doubtful. Should probably not be supported in its present form.

Poor: Proposal contains major scientific blunders and/or investigator is known to be incompetent. Should not be supported.

We recognize that the judgments involved may be partly subjective and there will customarily be other reviewers besides yourself.

B-30

NATIONAL SCIENCE FOUNDATION
PROPOSAL RATING SHEET

Reviewer

Proposal No.:
Investigator:
Institution:
Please return to:
If possible by:

Comments (*Continue on additional sheet if necessary*)

OVERALL RATING
☐ EXCELLENT
☐ VERY GOOD
☐ GOOD
☐ FAIR
☐ POOR

NSF Form 173, Jan 1976

Signature of Reviewer:

Other suggested reviewers (*optional*):

Verbatim but anonymous copies of reviews will be sent only to the principal investigator/project director on request. Subject to this NSF policy and applicable laws, including the Freedom of Information Act 5 USC 552, reviewers' comments will be given maximum protection from disclosure.

APPENDIX C

METHOD USED FOR ESTIMATING COMPONENTS OF VARIANCE

Estimation in the components of variance formulation of pages 28ff.

The form of the derivation is due to Jack Kiefer.

Suppose that proposal j has n_{ij} reviewers when it is reviewed by method i; let Y_{ij} be the average of these n_{ij} reviews. It is easily shown that

$$\sum_{j=1}^{50} \sum_{k=1}^{n_{ij}} (Y_{ijk} - Y_{ij})^2 \; / \sum_{j=1}^{50} (n_{ij} - 1)$$

is an unbiased estimate of $\sigma^2_{R,i}$. Next note that $Y_{2j} - Y_{1j}$, for any j, estimates $\alpha_2 - \alpha_1$; such estimates have variance $4\sigma_I^2 + v_j$, where

$$v_j = \sigma^2_{R,1} / n_{1j} \; + \; \sigma^2_{R,2} / n_{2j} .$$

Then

$$t = \frac{1}{50} \sum_{j=1}^{50} (Y_{2j} - Y_{1j}) = \hat{\alpha}_2 - \hat{\alpha}_1$$

estimates $\alpha_2 - \alpha_1$ with variance $4\sigma_I^2/50 + V/2500$, where

$$V = \sum_{j=1}^{50} v_j .$$

Using the unbiased estimates of $\sigma^2_{R,1}$ and $\sigma^2_{R,2}$ above, one can thus obtain an unbiased estimate of V; call the estimate $\hat{V}$. The quantity

$$\sum_{j=1}^{50} (Y_{2j} - Y_{1j} - t)^2$$

is an unbiased estimate of $49\,(V/50 + 4\sigma_I^2)$. Using $\hat{V}$, one can now get an unbiased estimate of σ_I^2. ($\hat{\sigma_I^2}$ and $\hat{V}$ yield an unbiased estimate of the variance of t.)

Now let

$$U = \frac{1}{50} \sum_{j=1}^{50} (Y_{2j} + Y_{1j}).$$

Then

$$\sum_{j=1}^{50} (Y_{2j} + Y_{1j} - U)^2$$

is an unbiased estimate of $49(V/50 + 4\sigma_p^2)$. Since there is an unbiased estimate of V, it is possible to obtain an unbiased estimate of σ_p^2.