Letters

NSF Peer Review (Continued)

There is an important gap in the evaluation of the peer review process in the National Science Foundation (NSF) reported by Cole, Cole, and Simon (20 Nov., p. 881). Estimates of random error by means of correlations are a function of the "range of talent" in the population, but the authors neglected this feature of their data. For example, if an institution of higher education accepts almost every student who applies for admission, that is, the institution has almost as many openings as applicants, the mean level of talent will be low and the variance large. In such a case, the correlation between two different but equally valid methods of measuring the talent of the applicants will be high. If another institution is highly selective, having available few openings for the number of potential applicants, there will be a great deal of self-selection in the decision to apply. The mean level of talent in the applicant pool will be high and the variance low. The random error in the units of measurement of talent may be the same in the two institutions, but the correlation between two equally valid measures will be substantially lower in the selective institution.

It is probable that the populations of research applications to NSF from which the authors drew their samples are like the applicants to the selective institution of higher education just described. In fact, the authors recognize this in their reference 7, in small type at the end of the article, but they do not describe the implications of self-selection and the resulting homogeneity of proposal quality for the size of correlations among raters. Had there been full discussion of these issues, the data would have been better understood and erroneous national publicity might have been avoided.

It may not be possible to quantify precisely the effect of restriction of range of talent in the present instance, but data pertaining to it can be obtained. How large is the pool of potential applicants in each of the several disciplines? Is there evidence that the quality of the people who apply to NSF is higher than the quality of those who apply elsewhere or who do not apply at all? If the NSF research budget were to be doubled next year (and other sources of funding remained constant), one would expect that the accuracy of funding decisions as measured by the correlation between two independent and equally valid assessments would increase substantially. A drastic reduction in NSF funding (while other funding remained constant) would reduce the accuracy of the funding decision measured in the same way. By a similar line of reasoning, the accuracy of NSF funding decisions was probably higher in 1969 than it is today.

It is ironic, and may even present Congress with a catch-22 situation, that reduced funding, though it may lead to greater care in review and to less measurement error, will lead to a lower correlation between ratings of proposals and thus to seemingly greater error.

LLOYD G. HUMPHREYS
Department of Psychology, University of Illinois at Urbana-Champaign, Champaign 61820

The conclusion that high variability in reviewers' evaluations of research goals gives rise to a significant element of chance (luck) in whether or not a proposal is funded may be valid for the system the COSPUP experiment tested. However, some programs at NSF use a much more intensive system. For example, all programs in the Division of Earth Sciences—where I was geochemistry program director from 1976 to 1978—use, in addition to mail peer review, a proposal review panel. The proposal review panel is composed of a group of scientific peers (five in the case of geochemistry), who provide an additional layer of judgment and selection on top of the mail reviews. This additional layer is, in my opinion, the most valuable part of the peer review system. Until a proposal is reviewed in a panel meeting the people who read and evaluate it are all acting as individuals: each mail reviewer reads the proposal in isolation, stares at the wall for a while, and records his or her judgment. Members of the review panel periodically receive copies of all proposals under consideration. All members read all proposals, with each member having primary responsibility for proposals appropriate to his or her subfield. The panel then meets with the program director, and the proposals are discussed and rated (on the same scale system used for mail reviewers) one by one. The panel members have available to them the opinions of all the (highly variable) mail reviewers, as well as each other's opinions, which develop by discussion.

It seems to me not hard to see that adding the panel review process to the mail peer review system should result in a much more rational selection of proposals for funding. I firmly believe that it does so, having observed NSF programs that operate without as well as with such panels.

I suggest that COSPUP test the full NSF peer review system. If the three programs whose review system was examined (chemical dynamics, economics, and solid-state physics) use panel review, COSPUP could select a second set of panels to complete the process, then compare the results with those of NSF. If the programs they chose do not (or did not) use panel review, the study should go back to square one, and programs representative of all NSF review systems should be tested. I have no doubt that the fuller process would prove to have far more repeatable outcomes. I also believe that it results in the funding of better science.

JOHN HOWER
Department of Geology, University of Illinois at Urbana-Champaign, Urbana 61801

The article by S. Cole *et al.* opens with the assertion that NSF "employs one form of the peer review system in making research grants." . . . The process described is in fact not the only form of peer review used by NSF. In biochemistry and biophysics, for example, this system of ad hoc reviewers is combined with review by a panel that contains expertise across the whole field. Where the reviews written independently by expert panel members and ad hoc reviewers are in substantial agreement, as is the case more often than not, funding decisions are relatively routine. It is when there are disagreements in the responses of experts that this system displays its great strength. The reasons behind the differing responses are addressed and debated, and a well-considered decision is reached that goes far beyond the blind averaging of scores. I believe that this procedure results in a much smaller chance component in mak-

ing funding decisions than does the use of either ad hoc reviewers alone or a panel alone, and probably represents the best selection system available. . . .

JOHN WESTLEY
Department of Biochemistry,
University of Chicago,
Chicago, Illinois 60637

. . . It is true, as the authors point out, that the present peer review system "compels" people to spend substantial amounts of time and energy composing proposals. Obviously most of the man-hours so spent are spent by authors of the middle three quintiles of proposals. Equally obviously, many hours are also so spent by "our most talented scientists." That even this is a "clear disadvantage . . . of the current peer review system" is not clear, however. As Dr. Johnson pointed out (in a rather different context), nothing so clears a man's mind as the knowledge that he is to be hanged on the morrow. It is easy to have an idea of what one wants to do without having a *clear* idea of what one wants to do; I suspect that wording proposals is often no mere exercise in composition but an important element in the process of giving one's ideas substance.

HENRY E. KYBURG, JR.
Department of Philosophy,
University of Rochester,
Rochester, New York 14627

With Atkinson (Letters, 18 Dec., p. 1292) and other members of the scientific community we regret that some members of the media have misread and incorrectly interpreted the results presented in our *Science* article. For instance, nowhere do we say that receipt of an NSF grant depends mostly on "luck." We did point out that getting an NSF grant was about half a result of the "luck of the reviewer draw" (in that we saw about 25 percent reversals rather than the 50 percent expected by chance alone), and we also pointed out that it was half a result of agreement among reviewers. This means that the current peer review system is decidedly superior to one based on random selection. Some of the critics of our article want to emphasize the 50 percent that is due to agreement rather than the 50 percent that is due to chance. But not one of them denies the validity of the main finding: that a new set of reviewers would produce a reversal rate of 25 to 30 percent and that reversals are largely a result of differences in the evaluation of these proposals by sets of apparently unbiased referees.

In discussing these findings with the press we emphasized that this study suggests the importance of more widespread funding of scientific research. If it is difficult to determine which project will lead to a major breakthrough, granting agencies should fund a wide range of research so as to reduce the probability that development of important ideas will be delayed because of lack of support. Lack of consensus is an inherent characteristic of science rather than of the NSF peer review system; thus we have never said or implied that the NSF was not efficiently and equitably run or that some other type of peer review system could overcome the problems resulting from lack of consensus.

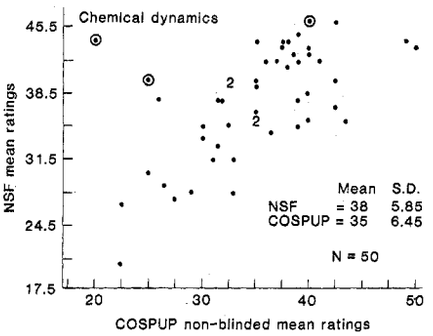

Fig. 1 (left). Scattergram of mean NSF ratings with mean COSPUP ratings of individual proposals. Encircled dots indicate proposals that received two or fewer reviews from either COSPUP or NSF reviewers. [Reproduced from (*1*), p. 28]

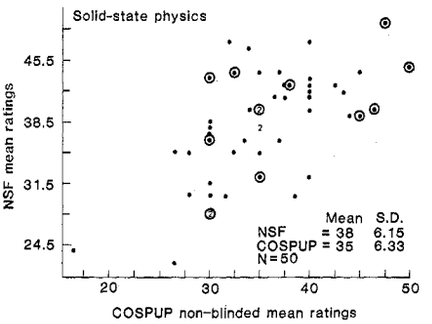

Fig. 2 (right). Scattergram of mean NSF ratings with mean COSPUP ratings of individual proposals. Encircled dots indicate proposals that received two or fewer reviews from either COSPUP or NSF reviewers. [Reproduced from (*2*), p. 30]

We are perplexed by Singer's statement (p. 1292) that our article "misrepresented" the results of the COSPUP report we coauthored with the committee (*1*). The conclusions reached in that report are virtually identical with those reached in our article. In his letter Singer selects one part of the findings that emphasize consensus (comparing COSPUP ratings with NSF decisions) and ignores those that don't (comparing COSPUP ratings with NSF ratings). Singer suggests that the proposals in the top quintile have only a small probability of being reversed, but in the report itself a concluding section written by the committee states: "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" (*1*, p. 58). It is important to recognize that a proposal rated in the top quintile by one group of NSF reviewers could have fallen into a lower quintile if rated by another group of reviewers. If the fact that the same proposal would either be funded or declined depending upon which group of reviewers was selected does not indicate that getting an NSF grant depends on the "luck of the reviewer draw," we would welcome another interpretation.

The standard deviations Cronbach (p. 1294) computes for the average of four reviews (3.8, 3.5, and 4.8) are far from small given the proposal standard deviations for the three fields (4.9, 4.9, and 7.6). Moreover, a substantial minority of proposals had three or fewer NSF reviewers. The scatterplot that Cronbach presents, showing mean ratings obtained by COSPUP reviewers against those obtained by NSF reviewers in the field of chemical dynamics (Fig. 1), illustrates a higher level of agreement between the two groups than figure 1 in our *Science* article, where proposal ranks are plotted. Chemical dynamics happens to be the field for which the plot of raw averages looks most pleasing. If Cronbach had chosen the identical scatterplot for the field of solid-state physics (Fig. 2), he might not have come to the same conclusion.

We agree with Cronbach that reviewer disagreement should not be disparaged as "random or nonrational." In fact, in the conclusion to our article we emphasize that our findings "should not be interpreted as meaning either that the entire process is random or that each individual reviewer is evaluating the proposal in a random way. . . . The great bulk of reviewer disagreement observed is probably a result of real and legitimate differences of opinion among experts about what good science is and should be." We thank Cronbach for pointing out the arithmetic error in our reference 9.

We will comment briefly on several other points made in the letters about our article:

1) Although Atkinson is correct in stating that the NSF decisions are not based upon numerical ratings in any technical sense, the Phase One report

(2), based upon 1200 proposals submitted to ten different programs, showed that an average of 67 percent of the variance in decision was explained by the mean rating. Of the 150 proposals used for Phase Two (*1*), on only 12 did the decision differ from the decision that would have been made by using mean ratings in a mechanical way.

2) Making verbatim reviews available and encouraging appeal of a decision the applicant believes to be unfair may be improvements in peer review, provided appeals receive equitable review. A study showing the number of appeals, the type of applicants who appeal, and the outcome of these appeals would be enlightening.

3) We strongly agree with Humphreys' point that the significance of our results must be interpreted in the light of self-selection of NSF applicants. (Space restrictions prevented us from expanding on our references 2 and 7, which dealt with self-selection.) It is possible that if we had asked a random sample of American scientists to write and submit proposals there would have been greater variance in the proposal means and a corresponding reduction in the ratio of reviewer variances to proposal variances. Therefore the relative significance of chance might have been reduced. However, this is not the situation NSF actually faces.

4) We agree with Hower and Westley that the use of panels improves the peer review process. The way in which panels work is discussed in the Phase One report (*2*). For one field included in the experiment, economics, NSF did use a panel. In that field the rate of reversals for COSPUP ratings compared with NSF mail ratings (28 percent) was very similar to that for COSPUP ratings compared with the NSF decisions (24 percent), decisions which were influenced by the panel. We note that the panel-augmented decision agrees strongly with the NSF mean ratings. The panel does reduce the reversal rate somewhat in the top quintile. There is no evidence yet that the reversal rate would have been lower if the COSPUP experiment had used either a substitute program director or a panel to make the decisions. A detailed examination of substantive comments made by reviewers for cases in which differences between the COSPUP and the NSF reviewers would have led to reversals suggests that the reversals were a result of legitimate intellectual differences rather than of "errors" by reviewers.

5) We agree with Kyburg that writing a proposal can be a very useful experience. However, it can also become an end in itself, resulting in a displacement of goals in which scientists spend almost as much time applying for funds as using them to produce new science.

We are pleased that Singer notes that the full report deals with questions other than funding reversals. COSPUP decided not to include in the Academy reports our analyses of additional topics which we believe shed light on peer review. These include a discussion of the effects of self-selection; data on peer appraisals of the reputations or "track records" of NSF applicants, and a comparison of consensus on reputations with consensus on proposals; and an analysis of pooled data on the probability of a reversal as a function of the number of reviewers as well as the variance structure of ratings of the proposals.

Finally, it should be noted that, although our experiment was based upon only 150 cases, the conclusions on consensus replicated those from the Phase One data on 1200 proposals (*3*). The variance structures of reviewer ratings in the ten fields studied were remarkably similar to the data produced by the experiment. Since reversals were found to be substantially explained by lack of reviewer agreement, we believe we would have found a similar reversal rate if the experiment had been replicated on the 1200 Phase One cases.

S. COLE

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

J. R. COLE

Center for the Social Sciences, Columbia University, New York 10027

G. A. SIMON

Department of Applied Mathematics, State University of New York, Stony Brook 11794

References

1. J. R. Cole, S. Cole, with the Committee on Science and Public Policy, *Peer Review in the National Science Foundation: Phase Two of a Study* (National Academy of Sciences, Washington, D.C., 1981).
2. S. Cole, L. Rubin, J. R. Cole, *Peer Review in the National Science Foundation: Phase One of a Study* (National Academy of Sciences, Washington, D.C., 1978).
3. J. R. Cole and S. Cole, *Nature (London)* **279**, 575 (1979).

The Curies' Nobel Prizes

In closing their informative description of Kai Siegbahn's research recognized by the 1981 Nobel Prize in physics (6 Nov., p. 629), Jack M. Hollander and David A. Shirley note that "the 1981 Nobel award to Kai Siegbahn is the fourth time that a father and son have both received the Nobel Prize." They conclude with mention of the Braggs, the Thomsons, and the Bohrs.

It is curious that the authors interested themselves with father-son Nobel laureates, rather than with the more general category of parent-child laureates. If they had considered the latter category, they would surely have included the mother-daughter and father-daughter awardees: Pierre and Marie Curie (1903 for radioactivity), Marie Curie (a second award in 1911 for the discovery of radium and polonium), and their daughter Irene Joliot-Curie (in 1935 for artificial radioactivity).

If we scientists are to claim that opportunities in science are open to women on a fair basis, as I have in the past, we must be careful to recognize women when they succeed.

DAVID EISENBERG

Department of Chemistry, Molecular Biology Institute, University of California, Los Angeles 90024

AAAS Election: Regression Toward the Mean

Last year we noted a change in the way the sexes fared in AAAS elections: The previous advantage that women enjoyed, once nominated, had greatly diminished (Letters, 6 Feb. 1981, p. 532). The 1981 election (21 Aug., p. 863; 4 Dec., p. 1115) shows a continuation of this pattern, albeit at a slower rate. Here are the percentages of those nominated who were actually elected, in contests with both sexes represented (omitting one grossly unbalanced race where 14 men were matched against a lone woman):

Year	Male (%)	Female (%)
1979	38.9	70.0
1980	44.6	58.8
1981	46.4	56.4

Thus a Tendency becomes a Trend. As properly cautious social scientists, we should, however, wait for a fourth year's data before advancing a Theory.

STEPHEN M. STIGLER
VIRGINIA L. STIGLER

5816 South Blackstone Avenue, Chicago, Illinois 60637

Erratum: In the article "Women scientists and engineers: Trends in participation," by Betty M. Vetter (18 Dec., p. 1313), a study by C. Rose was incorrectly cited in reference 12. The correct citation is C. Rose, *Academic Employment and Graduate Enrollment Pattern and Trends of Women in Science and Engineering* (Final Technical Report to the National Science Foundation, Evaluation and Training Institute, Los Angeles, Calif., 1978).