PEER REVIEW

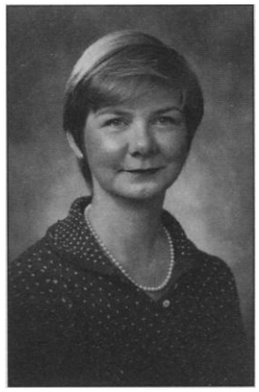
NIH Panel Urges Overhaul of The Rating System for Grants

Choosing winners from among the 30,000 biomedical scientists who seek federal grants each year is a delicate task, especially when more than three-quarters of grant proposals now get rejected. But tinkering with the peer-review process by which the National Institutes of Health (NIH) picks these winners is an even more delicate proposition. An internal NIH panel has, however, just issued a set of recommendations that could result in a fundamental change in the way peer-review panels weigh scientific merit.

NIH began an analysis of the way reviewers compute "priority scores" back in November 1994, when Wendy Baldwin, NIH's deputy director for extramural research, asked a group of staffers to consider how the system might be improved to focus reviews on the substance of research and sharpen critical judgments. After a contentious internal debate, the panel completed a draft report last winter; on 17 May, it submitted a final version to the advisory council of the NIH Division of Research Grants. The report makes 10 specific recommendations. The general aim, the report says, is to get reviewers to spread scores along a wide spectrum rather than clumping them at the "good" end, and to elicit clearer measures of the quality and significance of research ideas.

Last week, Baldwin released the 32-page report on the Internet and invited comment.* Baldwin and Constance Atwell of the National Institute of Neurological Disorders and Stroke, who chaired the umbrella group at NIH—the Committee on Improving Peer Review—under which these discussions took place, say they expect these recommendations will stir some anxiety and debate. They are eager to reassure NIH's constituents, however, that everyone will have a chance to comment before a final decision is made. "This is not a take-it-or-leave-it proposition. ... We don't have to accept all—or any—of the recommendations," Baldwin says. Already, one ad hoc member of the panel—biologist Keith Yamamoto of the University of California, San Francisco—so strongly disagrees with some of the panel's suggestions that he has written up and circulated an alternative set of recommendations.

The panel's most important recommendation, according to NIH staffers, is that peer reviewers should use a few explicit criteria in judging each application. At present, NIH asks reviewers to consider a list of a dozen criteria, but doesn't ask them to give specific comments on any. Reviewers simply vote a single, comprehensive score of 1 (excellent) to 5 (poor) in tenths of a unit. Instead, the panel proposes that grants be rated separately on each of three criteria: "significance," "approach," and "feasibility." Proposals would be given a score of 0 (poor) to 10 (excellent), in whole units, on each criterion.



Seeking input. Wendy Baldwin expects a debate.

The first criterion, Atwell says, covers the potential impact a project might have on its field. Reviewers would be asked to rate a proposal on the extent to which it could "make an original and important contribution to biomedical and/or behavioral science." While some people worry that this means grants would be rated according to program relevance, Atwell says that's not the intent. The second proposed criterion, "approach," would cover technical issues such as a project's design and methodology. The third criterion, "feasibility," would cover questions about the investigators' experience, preliminary data, ability to recruit subjects, staff, and other resource issues.

Hugh Stamper, extramural research director at the National Institute of Mental Health, says that psychometric studies show that reviewers give more reliable, reproducible ratings if they are asked to "disaggregate" the elements of a decision. (Stamper co-chairs the panel that authored this report, the Committee on Rating Grant Applications, a subcommittee of Atwell's group.) For one thing, reviewers wouldn't be able to "slip private criteria" into their judgments. Atwell also notes that reviewers often fudge the most important question of all—whether the proposed research is significant. Opinions on this key point "get mushed up and blurred," she says, "or don't get mentioned at all."

On the need for better criteria, Yamamoto agrees. "Right now," he says, "we often read grants in which the [scientific] impact of the proposed research is imperceptible. The tendency of reviewers is not to say, 'This grant is boring.' Instead, we write several pages describing technical flaws." The applicant may respond by going back to the lab, making technical fixes, and submitting the boring idea once again, according to Yamamoto. Such "amended" applications now clog the system, and NIH's reformers would like to be able to deliver a clear message to grant applicants that technical fixes will not convert a dull project into an interesting one.

Yamamoto would also like to send an additional message by including a fourth criterion: "creativity or innovation." He argues that it is important, particularly when money is tight and reviewers are betting on proposals that look like sure winners, to include an incentive for risk-taking. Otherwise, Yamamoto fears, applicants will not dare to seek funding for their best ideas. But members of the NIH panel didn't agree. They left innovation out, Stamper says, because it seemed a bad idea to suggest that every grant should strive for creativity. That topic, Stamper thinks, could be included in the term "significance."

One of the most contentious issues, Stamper notes, is whether or not reviewers should combine the results of these specific ratings into a single score. After going back and forth for months, the panel decided not to endorse a single score. Yamamoto disagrees strongly with this recommendation. "It would be very unfortunate," he says, "not to arrive at an overall merit score. That would mean the study section would fail to take responsibility" for making the final appraisal, deferring instead to NIH. The panel did hedge, however, by recommending that if NIH wants a single score, it should calculate the unweighted average of the scores for the three criteria.

Yamamoto also takes issue with a recommendation that all scores be standardized to the performance of the individual reviewer. At present, NIH ranks scores according to records established for each peer group. The NIH staff has proposed switching to a system in which NIH would keep track of each individual's scoring style (similar to a golf handicap) and adjust results to reflect the individual's grading habits, using a technique not spelled out in the report. This, Yamamoto says, would be "a very bad idea" for the simple reason that "individuals do not display standard behavior." Fine-tuning the system to this degree, he thinks, would create distortions.

All of these issues, Baldwin says, will be discussed at institute council meetings and in public forums with the NIH constituency before any decisions are made. Baldwin says the system "is not going to change overnight," although working groups are expected to come up with an implementation plan by 1 October.

–Eliot Marshall

* "Report of the Committee on Rating Grant Applications," co-chaired by Hugh Stamper and Walter Stolz. The report can be obtained at http://www.nih.gov/grants/rga.htm.