

## On My Mind: Enhancing Peer Review at NIH...What Users Think About The Changes

A belated Happy New Year. I hope that the coming year is a good one for you...a year of interesting and successful research, many grants and personal fulfillment.

Recently, a client and I had a conversation about the changes in peer review that have recently been initiated at NIH. During the conversation, the client said, "Well, Norm, what do you think about them?" I replied that what I thought was less important than what the users thought. Since I was involved in the development of 'user' surveys conducted by NIH as part of the continuous review of peer review, I thought that it would be a good idea for me to share some of the findings.

During our conversation I also pointed out that one of the characteristics shared by successful applicants is that they learn as much as they possibly can about the assessment and funding decision process used by the agency to which they are applying. This is true whether or not they are applying to a federal agency or a private foundation. As a corollary, they also keep informed about any changes that take place in that process.

As most applicants and potential applicants for NIH funding are now keenly aware, the peer review process has undergone an extensive revamping under the title "Enhancing Peer Review (<http://enhancing-peer-review.nih.gov/>)." Briefly, the changes in peer review at NIH introduced in May 2009 included:

- Shifting from a 1-5, 41 point, scale that used single word adjectives (i.e., outstanding, excellent, average, poor) to define the quantitative scale, to a 9-point scale that provides well defined anchor points giving concrete meaning to each of the numeric points on the scale
- Providing applicants with numeric scores for each review criterion that receives a score from all assigned reviewers. These scores are included in the summary statement of all applications, even those that are not discussed during the review meeting as a way of providing applicants with additional information in the form of a quantitative assessment of their proposed research.
- Employing a structured format for written comments in bullet format. This was introduced as a way of reducing the burden on reviewers while, at the same time, providing applicants with the key points that influence the quality of the proposed research.
- Enhancing the five core review criteria, used previously, with an additional one aimed at providing an assessment of the impact of the research (i.e., the likelihood that the outcome of the research will "exert a sustained, powerful influence on the research field")
- Clustering the review of applications proposing clinical research
- Clustering the review of applications submitted by New Investigators

Since the introduction of these changes, additional ones have been adopted. These include changes in the length and format of the grant application ([http://enhancing-peer-review.nih.gov/page\\_limits.html](http://enhancing-peer-review.nih.gov/page_limits.html)). We won't be discussing these in this article.

## Continuous Review of Peer Review: Stakeholder Surveys

One of the elements of the initiative is termed “continuous review of peer review.” According to the NIH web site “Enhancing Peer Review At NIH”,

“...a critical component in assuring the core values of peer review is the continuous review of peer review activities. NIH has established a Peer Review Evaluation Group to lay the foundation for continuous review of peer review and has commenced a dynamic effort to assess the cumulative outcomes of the concurrent changes being brought about by the peer review enhancements. The group’s efforts include:

- On-line surveys of multiple audiences.
- Data-driven mechanisms to evaluate review outcomes.
- Peer review pilots and assessment of those pilots...”

And, indeed, one of the very first activities under this rubric was a survey of five groups of stakeholders...reviewers, applicants, members of national advisory committees to the various ICs, NIH scientific review officers (SROs), and NIH scientific program, officers (POs)...focusing on the initial changes in peer review. We are very pleased to have been given the opportunity to participate in this activity as part of the Peer Review Enhancement Team.

### Stakeholder Surveys: Background

Links to the surveys are available on the Enhancing Peer Review web site. Applicants and reviewers were surveyed in December 2009 shortly after the changes were introduced. Only one complete application cycle had occurred when these surveys were deployed. NIH had intended to obtain a pre-change measure of individuals’ feelings about peer review as a baseline. This was to be followed by a post-change survey in order to gauge the impact of the changes on acceptability. However, the changes in peer review were introduced before all stakeholders could be surveyed when NIH reviewed applications submitted in response to the government-wide ARRA initiative.

As a result it was not possible to obtain baseline input about how reviewers and applicants felt about the old peer review systems, so they were asked to respond to the survey **based on their most recent experience**...that is to say, either with the old peer review system for those who hadn’t submitted and/or reviewed an application to an ARRA-based RFA (Applicants = 374\*; Reviewers = 221) or the revised system used beginning in May, 2009 for the ARRA applications (Applicants = 504; Reviewers = 537). Surveys for members of national advisory councils (N=291) were conducted during the January 2010 council round, while those for SROs (N=288) and POs (N=437) were administered in April 2010 after two full application cycles had occurred well after many of the changes had been introduced. [*\*Numbers in parentheses are the number of individuals providing usable surveys.*]

Results of the surveys have recently been posted on the NIH Enhancing Peer Review web page ([http://enhancing-peer-review.nih.gov/docs/Enhancing\\_Peer\\_Review\\_Report.pdf](http://enhancing-peer-review.nih.gov/docs/Enhancing_Peer_Review_Report.pdf)). According to the report applicant and reviewer survey responses were analyzed using standard statistical methods developed specifically for categorical data (e.g., strongly agree, agree, do not agree). No information is given about the analytic method used for SROs, POs or council members, although one can assume that conclusions were drawn based on direct comparisons of frequencies of responses to the various

categories. Also, in reading the report, it is apparent that not all results are reported. No explanation is given for inclusion or exclusion of findings.

It is stated in the report that demographic and experience factors of respondents (e.g., age, gender, funding history, prior experience in NIH peer review) “...were found to be statistically significant for a small number of survey questions, and there was no compelling pattern to these results that warranted additional analyses or hypothesis testing beyond the original focus of the surveys.... (p. 7). Unfortunately, details of the actual survey questions or the pattern for which the factors were statistically important are not provided so we were not able to see for ourselves whether the statement was, in fact, justified. This was also true in reporting some of the other survey findings.

### **Stakeholder Surveys: Findings**

The report is organized around the main points of the survey, corresponding to the changes introduced in the peer review system. I have tried to extract information from the survey that’s of interest to applicants. The conclusions stated in this paper are not simply a reiteration of those presented in the report. They are mine and sometimes do not coincide with and/or go beyond those presented in the report.

- **9-point scoring scale with descriptive anchor points**

The previous rating scale, nominally a five-point scale, was in reality was a 41-point scale because assigned scores in 0.1 increments. The change to a 1-9 integer scale was introduced because it was believed that it was not possible to discriminate between scientific merit in 0.1 increments, particularly since the 1-5 scale was only defined in single word adjectives such as ‘outstanding’, ‘excellent’, ‘good’ or ‘poor.’

The survey asked, among other things, whether there was sufficient range in the 9-point scale for reviewers to meaningfully distinguish among applications of varying scientific merit. Judging from the data reported, it appears that the scale is adequate for the vast majority. What isn’t included in the survey or the report is the extent to which the change in descriptive anchor points plays a role in this finding. While I suspect that the enhanced description of what a given numeric score means is very important in the acceptability to reviewers of the change, from an applicant’s perspective, ***the important point is that reviewers feel comfortable with the new scale and it is now important for applicants to understand the scale in order to know how their applications are being rated.***

There was also interest on the part of NIH whether or not members of national advisory councils would adapt quickly to the change in the scoring system. Results of the survey show that approximately 2/3 of the respondents agreed or strongly agreed that they were able to understand the translation. Not reported in the document are findings about the new scoring system from the applicants, POs or SROs.

- **Use of criterion scores**

Changes in the review process involved asking each reviewer and discussant assigned to an application to provide a separate score for each of the five review criteria (i.e., Significance, Investigators, Innovation, Approach, and Environment) prior to the review meeting. This change was introduced for several reasons including to facilitate the discussion of applications among reviewers at the review meeting as well as to provide additional feedback to applicants since the scores are included

in the summary statement of all applications, even those that are not discussed during the review meeting.

Results of the survey indicate that slightly more (41%) applicants found the scores helpful than those that didn't (34%). About the same percentages (45% vs. 34%) of applicants reported that the criterion scores were helpful in focusing them on problem areas that could be corrected and presumably improve the scientific merit of their application. About twice as many reviewers (i.e., 49% vs 25%) tended to view the criterion scores as being helpful in communicating why applications were not discussed at the review meeting.

Selecting from one of eight options reflecting the changes in peer review as to which was the most helpful in advising applicants (i.e., 9-point scale, enhanced review criteria, reviewer scores for each criterion, use of an overall impact score, use of critique template, bulleted comments, clustering of applications from new investigators), 45% of POs surveyed indicated that "**reviewer score for each of the review criteria**" was most helpful in advising applicants. However, it is also important to note that 22% of the surveyed POs indicated that none of the changes in peer review were helpful.

- **Consistency of criterion scores with overall impact scores**

Reviewers are instructed to assign an overall impact score to applications that are discussed at the review meeting. This score represents an assessment of the impact of the research as indicated by likelihood that the outcome of the research will "**exert a sustained, powerful influence on the research field.**" In other words, if all of the technical issues identified for an application are successfully addressed, reviewers are asked to predict how important the proposed research will be to the field in general. The overall impact score is said to be independent of the five criterion scores (i.e., Significance, Investigators, Innovation, Approach, and Environment) and reviewers are instructed not to simply assign an average of the criterion scores as the overall impact score.

Both the SROs (43%) and POs (45%) tend to disagree with the statement that the overall impact score is consistent with the individual criterion scores. It's not clear from the report whether this means that the overall impact scores are being treated independently from the individual criterion scores by reviewers (i.e., that reviewers are following instructions about the various score components) or that SROs and POs expectation that the overall impact scores being in line with the individual criterion scores was not met. There are other data available that we will discuss in a subsequent **What's On My Mind** article that address this issue directly. No results are reported for reviewers, applicants or members of advisory councils.

- **Bulleted Comments and Structured Critique Template**

The use of the bulleted comments and structured critique template for written reviews was introduced for several reasons. On the one hand, using bulleted comments and a structured/guided critique template was viewed as a way of reducing effort in providing written assessments as well as focusing discussions and hence shortening review meetings. In addition, bulleted comments and structured critiques were also thought to provide applicants with less ambiguous and key points that influenced the reviewers' assessments of the quality of the proposed research.

The report focuses on three aspects of the bulleted/structured format:

- whether or not the bulleted comments helped focus on strengths and weaknesses

- whether the structured templates helped to streamline the preparation of critiques, and
- whether the bulleted comments provided the information needed to understand why the applications were not discussed.

Applicants saw no differences between the narrative and bulleted format in helping them **identify strengths and weaknesses**. Approximately two-thirds indicated that the narrative form was helpful and the same percentage said that the bulleted format was helpful. About 45% of the SROs indicated that the bulleted critique format was helpful in focusing on factors that influenced score while 31% of their colleagues disagreed with it.

Reviewers were almost unanimous (93%) in their view that the narrative format was adequate for capturing **strengths and weaknesses** while significantly fewer (67%) of them agreed with the same statement in reference to the bulleted format. Their view about the bulleted critiques is consistent with the view expressed by more than one-half of the POs that the new summary statements were not helpful in explaining the recommendations of the review group. However, 66% of reviewers who had experience only with the narrative format indicated that the narrative **format helped them complete their reviews efficiently** while 76% of reviewers who had experience with both formats indicated that bulleted/structured format allowed them to complete their reviews efficiently.

Applicants' agreement with the ability of the new summary statements to identify strengths and weaknesses did not carry over to their view of whether these summary statements **explain why their application was not discussed** by the review group. In fact, there was agreement among all stakeholders that the new format for summary statements were not able to explain why an application wasn't discussed.

- **Enhanced review criteria**

Enhancing the review criteria by defining scientific merit as overall impact of the research, independent of how an application measures up to the traditional five criteria, was viewed similarly by SROs and POs. Approximately, one-third viewed this move as providing greater clarity of the strengths and weaknesses of the application, one-third view it as not providing greater clarity and one-third believe that it doesn't make a difference. No other data are provided in the report about reviewers', council members', and applicants' views of this change nor was there any fine-grained analysis of the responses by SROs and POs.

- **Clustering of applications from new investigators or involving clinical research**

Approximately 70% of reviewers agreed or strongly agreed that clustering applications from new investigators resulted in a more consistent review while only 54% felt the same about clustering clinical research. Further, both POs (72%) and SROs (47%) selected most often the statement that clustering applications from new investigators was a positive contribution of the Enhancing Peer Review initiative.

- **Overall Satisfaction: Applicants and Reviewers**

Applicants that applied after the review changes were instituted exhibited no preference (old-38% vs new – 39%) for either the old or new review system, while there was a clear preference for the new system (old – 25% vs new – 60%) among reviewers who participated in reviews after the new system had been introduced. Both applicants and reviewers expressed overall satisfaction with and

fairness of the new system in comparison to the old system, although in both cases there tended to be greater satisfaction with the old system, a finding that may simply reflect a normal response to change.

- **Overall Satisfaction: SROs, POs and Advisory Council Members**

Overall satisfaction with the new system among SROs, POs and members of advisory councils was quite different. SROs were evenly divided in their preference for the new (35%) and old (33%) systems while 31% had no preference. Members of advisory councils showed a preference for the new system (45%) over the old (30%) with only 24% expressing no preference. POs, in contrast, showed a very strong preference for the old (45%) over the new (27%) with 24% indicating no preference. Consistent with their preference for the old system, slightly more POs indicated an overall dissatisfaction (43%) than satisfaction (40%) with new peer review process. SROs tended to show greater overall satisfaction (46%) than dissatisfaction (35%) with the new system.

At the same time, POs and SROs tended to rate the new system as fair, although slightly more SROs were more favorable in their fairness ratings and slightly more POs indicated that the new system was unfair.

### **Next Steps**

It appears that NIH is planning to continue to monitor stakeholders' responses to modifications in the peer review system. As stated at the end of the report:

“NIH has been continuously monitoring feedback on (*sic*) the peer review changes since their inception and has already implemented a number of refinements to the peer review system in response to this information. Most notably, in January 2010, additional guidance was issued to clarify distinctions between Overall Impact and the reviewer criterion Significance; in September 2010 NIH began requiring reviewers to include a narrative statement to explain the overall impact score. NIH will reassess stakeholder opinions of the scoring system, the critique format and other peer review processes at a later date. In addition, NIH's ongoing review of the NIH peer review system will also examine the shortened applications and alignment of research plans with the NIH review criteria.”

That refinements may continue makes it all the more important that grantees, applicants and potential applicants continue to monitor the NIH website in order to remain current with any of these changes. In addition, it appears that 'continuous review of peer review' means more than monitoring stakeholder perceptions. Several quantitative studies of reviewer behavior in assigning scores to the various criteria have been conducted. The results of these studies are very important in shaping how applicants approach the development and revision of their applications. It will be the topic of the next **'On My Mind'** article.

Good luck and keep in mind that successful applicants are well versed in the review and funding processes used by the source of the support for their research.

Norm Braveman  
Braveman BioMed Consultants  
Rockville, MD  
January 10, 2011