NATIONAL INSTITUTES OF HEALTH

2007-2008 PEER REVIEW SELF-STUDY

FINAL DRAFT

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The National Institutes of Health (NIH) has a longstanding history of supporting the most promising and meritorious biomedical and behavioral research using a broad range of approaches, strategies and mechanisms. The NIH peer review system has been adopted internationally as the best guarantor of scientific independence. However, the increasing breadth, complexity, and interdisciplinary nature of modern research have created challenges for the system used by the NIH to support biomedical and behavioral research and for peer review, the cornerstone of the research enterprise. Thus, the NIH recognizes that as the scientific and public health landscape continue to evolve, it is critical that the processes used to support science are fair, efficient, and effective.

The NIH 2007-2008 peer review self-study consists of two, discrete phases: a diagnostic phase and an implementation phase. The goal of the first, diagnostic phase, reported herein, was to identify the most significant challenges to the system used by the NIH to support science and propose recommendations that would enhance this system in the most transformative manner. Specific implementation issues were purposefully not considered during this phase of the project; these details will be articulated during the second, implementation phase.

The diagnostic phase of the peer review self-study led the NIH and its stakeholders to articulate seven major challenges and associated goals and recommended actions to address each of them. The analysis clearly revealed the "systems" nature of peer review, pointing to the need for an integrated approach to enhancement. While each recommended action must be evaluated according to its own merit, the most optimal enhancement to the system will be achieved through the synergistic effects of multiple proposed actions.

Above all, it is critical that the NIH maintain the core values of peer review: scientific competence, fairness, timeliness, and integrity. When striving to fund the "best" science, the NIH must consider many factors, including scientific quality, public health impact, the mission of an NIH Institute or Center, and the current NIH portfolio.

Challenge 1: Reducing Administrative Burden on Applicants, Reviewers, and NIH Staff

For many investigators, staying funded is a time- and labor-intensive exercise that can compromise the practice of research.

Goal: To reduce the number of applications that need to be submitted by helping applicants make faster, more informed decisions to either refine an existing application or develop a new idea

Recommended Action:

- Provide unambiguous feedback to all applicants by establishing a "Not Recommended for Resubmission" (NRR) category and by providing scores for all applications.

Goal: To focus on the merit of the science presented in the application and not the potential improvements that may be realized following additional rounds of review

Recommended Actions:

- Eliminate the "special status" of amended applications by considering all applications as being new.
- Shorten summary statements by focusing solely on the merit of the science as presented.

Goal: To reduce application length to focus on impact and uniqueness/originality, placing less emphasis on standard methodological details

Recommended Action:

- Shorten the length of the application and align it to specific review elements.

Challenge 2: Enhancing the Rating System

The rating system that informs NIH peer review is central to every activity, and thus it is critical that the NIH carefully consider carefully ways to ensure that rating is both as accurate and informational as possible for both applicants and the NIH.

Goal: To focus and elevate the level of discourse of the study section
Goal: To provide unambiguous feedback to applicants
Goal: To enhance the consistency of rating and to engage all charter review members in the review of each application

Recommended Actions:

- Modify the rating system to include scores and ranking.
- Rate multiple, explicit criteria individually, but provide an independent overall score and ranking.
- Provide unambiguous feedback to all applicants by establishing a "Not Recommended for Resubmission" category and by providing scores for all applications.
- Restructure the application to reflect the rating criteria.

Challenge 3: Enhancing Review and Reviewer Quality

The cornerstone to review quality is recruiting and retaining excellent reviewers. Thus, improving review quality means addressing the larger problem of changing the culture of review.

Goal: To enhance review quality

Recommended Actions:

- Engage more reviewers per application
- Pilot the use of "prebuttals" for applicants and/or reviewers to correct factual errors in review
- Pilot anonymous review in the context of a two-level review system
- Enhance reviewer, study section, and scientific review officer training

Goal: To enhance reviewer quality

Recommended Actions:

- Create incentives for reviewers, including more flexible service and flexible deadlines for reviewer grant submissions
- Link potential review service to the most prestigious NIH awards
- Analyze patterns of participation by clinician scientists in peer review and provide more flexibility to ensure their continued involvement in review
- Continue piloting the use of patients and/or their advocates in clinical research review

Goal: To ensure the best use of charter review member time and expertise

Recommended Actions:

- Shorten application and summary statement length
- Have charter review members explicitly rank applications

Challenge 4: Optimizing Support for Different Career Stages and Types

Supporting early-career investigators emerged as a top challenge during the diagnostic phase of the 2007-2008 peer review self-study, and it has been the top priority of the NIH leadership for many years. However, there is also a need to enable greater productivity of highly accomplished NIH investigators, with less administrative burden to applicants and reviewers.

Goal: Early-career investigators should at a minimum be on par with established principal investigators in application success rates.

Recommended Actions:

- Continue to fund more R01s for early-career investigators
- Pilot the ranking of early-career investigators against each other
- Pilot the review of early-career investigators separately by generalists, to enhance risk-taking and innovation or uniqueness by applicants
- Take into account investigator/institutional commitment criteria for early-career investigator review

Goal: To enable greater productivity of highly accomplished NIH investigators, with less administrative burden to applicants and reviewers

Recommended Action:

- Refine the NIH MERIT/Javits/NIH Director's Pioneer Awards and, perhaps, other mechanisms to enhance productivity of the most accomplished investigators and to add to the pool of accomplished investigators available as potential reviewers.

Challenge 5: Optimizing Support for Different Types and Approaches of Science

Diverse types of science are needed to fulfill the NIH's mission to improve the nation's health, and peer review must accommodate the NIH's need to strike an appropriate balance among these.

Goal: To provide clear opportunities for applications proposing transformative research

Recommended Action:

- Use the NIH Director's Pioneer, NIH Director's New Innovator, and the Exceptional, Unconventional Research Enabling Knowledge Acceleration (EUREKA) Award programs as starting points to develop a path to invite, identify, and support transformative research, expanding the number of awards to a minimum of 1 percent of all R01-like awards.

Goal: To ensure optimal review of clinical research

Recommended Action:

- Determine the underlying causes of clinical research application submission patterns and results in the Center for Scientific Review (CSR) and NIH Institute and Center (IC) panels and consider corrective actions if needed.
- Ensure participation of adequate numbers of clinician scientists by providing more flexible options for review service.

Goal: To ensure optimal review and support for interdisciplinary research

Recommended Actions:

- Analyze applications that are interdisciplinary in nature with respect to referral patterns for review, assignment for secondary review and funding consideration, and success rate.
- Employ an editorial board model for the review of interdisciplinary research.

Challenge 6: Reducing the Stress on the Support System of Science

Regardless of the numerous and complex issues that stress the system used to support U.S. biomedical and behavioral research, resources will always be finite in nature. The NIH must continue to guide the distribution of these resources through careful and transparent prioritization in concert with the NIH's stakeholders.

Goal: To ensure the optimal use of NIH resources

Recommended Actions:

- Require, in general, a minimum percent effort for investigators on research project grants.
- Analyze the incentives inherent in the NIH system of funding that have been driving the rapid expansion of the U.S. biomedical research system in recent years and explore with stakeholders whether these incentives should be reduced or eliminated.
- Analyze the NIH contribution to the optimal biomedical workforce needs.

Challenge 7: Meeting the Need for Continuous Review of Peer Review

Finally, it is critical that the NIH establish data-driven mechanisms to evaluate review outcomes and to assess the success of pilot programs. This effort must be highly dynamic, to match concurrent changing landscape of biomedicine.

Goal: To assure the core values of peer review

Recommended Actions:

- Mandate a periodic, data-driven, NIH-wide assessment of the peer review process.
- Capture appropriate current baseline data and develop new metrics to track key elements of the peer review system.

Additional detail and data supporting each of these recommended actions is presented in the *Challenges, Goals, and Recommended Actions* section of this document.

A CONCISE HISTORY OF PEER REVIEW

In broad terms, peer review has an expansive history (1). The first documented description of a peer review process has been reported to be that described more than a thousand years ago in the book *Ethics of the Physician*, authored by Syrian physician Ishaq bin Ali al-Rahwi ($_{CE}$ 854-931). This work outlines a process whereby a local medical council reviewed and analyzed a physician's notes on patient care, to assess adherence to required standards of medical care (2,3). However, it was not until the 1600s, nearly two centuries after the invention of the printing press, that the first scientific journal, *Philosophical Transactions*, appeared. The introduction of this publication marked the first instance of an editor making decisions about what to publish. A century later, members of the Royal Society of London took over the editorial responsibility of the journal, having a select group of people review manuscripts for publication.

In the mid-1900s, the increasing diversity and specialization of science created the need for seeking outside assistance and the recruitment of outside reviewers. In 1944, re-codification of Public Health Service (PHS) laws into Public Law 410 included Section 301 to provide the PHS with authorization for research grants, thus giving the PHS overall authority that had in 1938 been restricted to the National Cancer Institute (NCI). Dr. Cassius Van Slyke was assigned to the NIH to direct this activity, and together with Dr. Ernest Allen, established the Office of Research Grants. Within a few months, this entity became the Division of Research Grants (DRG) (see Figure 1 for a timeline of NIH policies and actions related to peer review).

It was decided that the research grant mechanism would be used for support as opposed to contracts, due to the concern that grantees should not be burdened with contract requirements existing at that time, such as quarterly financial and scientific progress reports (4). Study sections consisted of non-governmental scientists had responsibility for the scientific evaluation of all research grant applications, leading to the two-level review system of today. These study sections reviewed applications based on scientific merit as well as confidence in the principal investigator. The National Advisory Health Council and the councils of the respective NIH Institutes and Centers (ICs) relied heavily on the recommendations of the study sections. However, as is the case today, they also considered program objectives and public health need. Scientists selected to serve on DRG study sections were leaders in their fields; many were Nobel laureates and Lasker awardees. Initially, these scientists reviewed individual applications, while also assessing the state of the science in their own fields of research, a role later assumed by IC staff and councils.

When NCI began receiving a separate appropriation in the mid-1940s the NIH appropriation contained funds for all other NIH research grants. As several ICs were established, the grants and funds represented by the categorical interests were transferred from DRG to the respective ICs and became the research-grant base for each IC's separate appropriation. In the early years, the DRG Chief had the authority, later transferred to the Associate Director of NIH, to issue new or modified policies. With

consultative advice of study section and council members, these officials met with appropriate IC representatives, obtained approval from the NIH Director, and issued policies without further clearance, sometimes within weeks. In today's world, the introduction of new policies takes months and sometimes years. Indeed, this is necessary due to the increased size and complexity of the NIH.

The most recent assessment of the NIH peer review system occurred in 2000, when the NIH-commissioned Panel on Scientific Boundaries for Review published the findings of its comprehensive examination of the organization and function of the CSR review process. This "Boundaries Report" (5) proposed two implementation phases. Phase I derived a revised set of Integrated Review Groups (IRGs) and outlined cultural norms to govern the CSR review process. Phase II established the scientifically related study sections that populate each IRG on the basis of principles outlined in the report.

More recent proposals to address peer review have suggested methods to ensure that reviewers welcome innovation (6,7,8,9), properly evaluate clinical research (10), and reduce logistical burdens (6,11). Because implementing these ideas on a broad scale would likely have intended and unintended consequences, the NIH recognizes the need to carefully consider the outcomes on all participants in the peer review process: the scientific community, the public, and the NIH (6,12,13,14,15,16).

In 2007-2008, leaders from across the scientific and public communities joined forces with the NIH to examine the current peer review system and consider potential ways to optimize it. Above all, the NIH wants to ensure that the agency will be able to continue to meet the needs of the research community and the public. The NIH peer review 2007-2008 self-study that is outlined in this report was co-led by an external working group (ACD WG[1]) co-chaired by Dr. Keith Yamamoto of the University of California, San Francisco, and National Institute of Dental and Craniofacial Research Director Dr. Lawrence Tabak and an internal working group (SC WG[2]) co-chaired by Dr. Tabak and National Institute of General Medical Sciences Director Dr. Jeremy Berg.

[1]The Advisory Council to the Director Working Group on Peer Review: http://enhancing-peer-review.nih.gov/rosters/acd.html

[2] The Steering Committee Working Group on Peer Review. http://enhancing-peer-review.nih.gov/rosters/adhoc.html

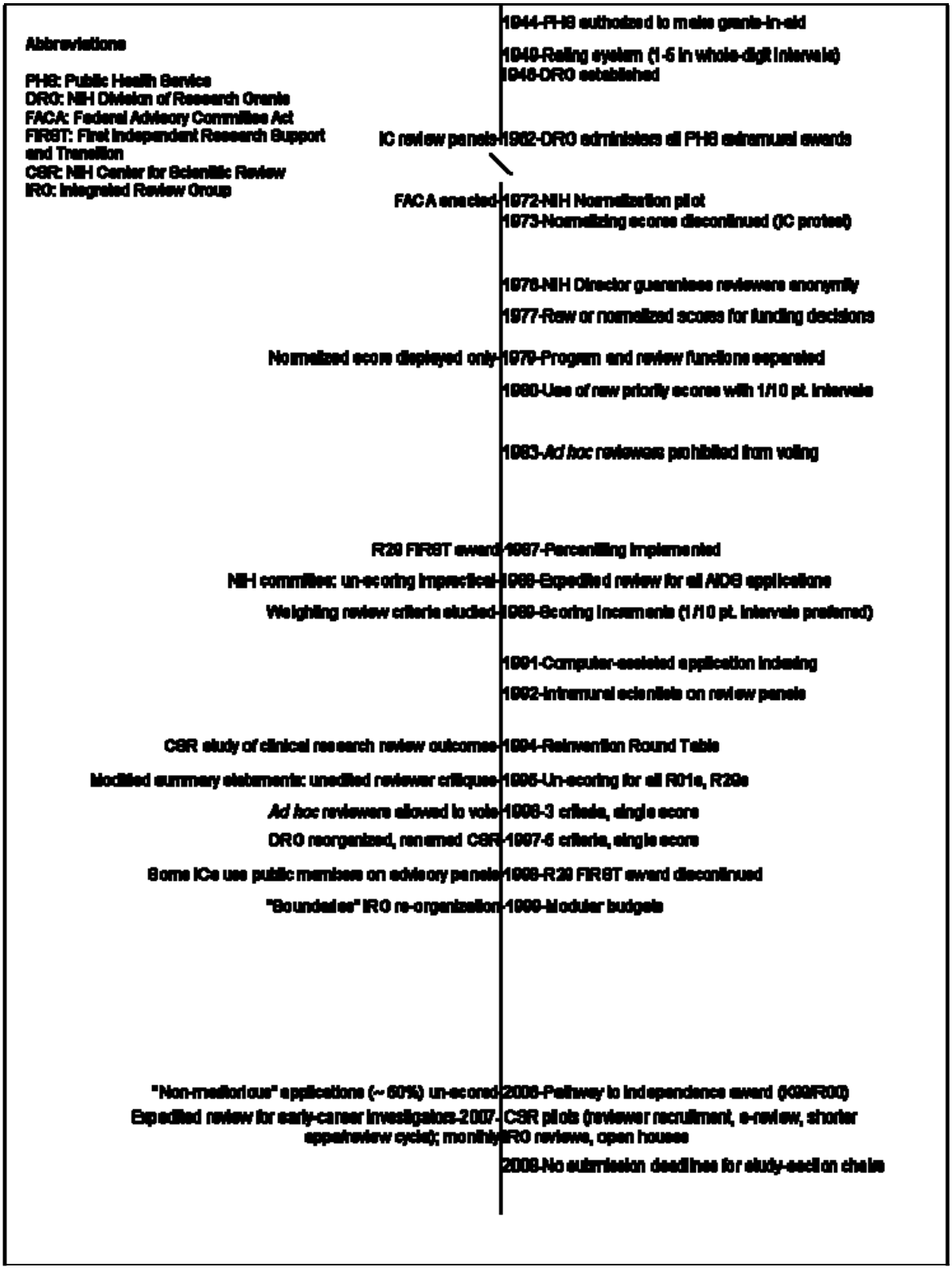


Figure 1. Timeline of NIH peer review activities and policies

PEER REVIEW AND THE NIH FUNDING PROCESS

The peer review process as it relates to NIH funding is a multifaceted, multi-stage endeavor. The NIH makes funding decisions using a range of criteria, to balance several issues including scientific quality, potential impact, portfolio balance, and relevance to the NIH mission.

The objective of the first stage of peer review is to evaluate and rate the scientific and technical merit of proposed research or research training. This takes place in Integrated Review Groups (IRGs, or "study sections") organized and managed by CSR. In addition, Review Branches of the ICs manage their own scientific review groups that evaluate applications submitted in response to special solicitations such as Requests for Applications, and for unique programs. The NIH Office of Extramural Research (OER) is integral to the overall process, since it manages the development and implementation of peer review policies and procedures across the NIH.

For most research grant proposals, after consideration and discussion, study sections assign the application under review a single, global score that reflects its overall impact relevant to significance, approach, innovation, investigator, and research environment. In this scheme, the best possible priority score is 100 and the worst is 500. Individual reviewers mark scores to two significant figures, and the individual scores are then averaged and multiplied by 100 to yield a single overall score for each application. Most research grant applications are then given a percentile rank, based on scores assigned to applications reviewed during the current plus past two review rounds.

Percentiles help indicate the spread of applications within a study section review. A percentile roughly translates to the percentage of applications receiving a better priority score during a one-year interval. The NIH uses percentiling to address the potential problem of reviewers giving applications better priority scores to the point where the scores had little meaning. Percentiles counter this trend by ranking applications relative to others scored by the same study section. However, even with percentiling, priority scores usually cluster in the "outstanding" range. Historically, reviewers have typically given as many as two-thirds of their applications priority scores between 100 and 200. The NIH includes unscored applications in the percentile calculation. Since the number of unscored applications varies by study section, including them affects the percentile distribution and makes percentiling fair across study sections.

In the second stage of peer review, the NIH National Advisory Councils consider the results of the first stage of peer review and make recommendations to ICs (Figure 2). Composed of scientists from the extramural research community and public representatives, advisory councils ensure that the NIH receives advice from a cross-section of the U.S. population in the process of its deliberation and decisions about funding. Councils meet three times per year, coincident with NIH application submission deadlines. Ultimately, the final funding decision for all submitted applications rests with the director of the funding IC.

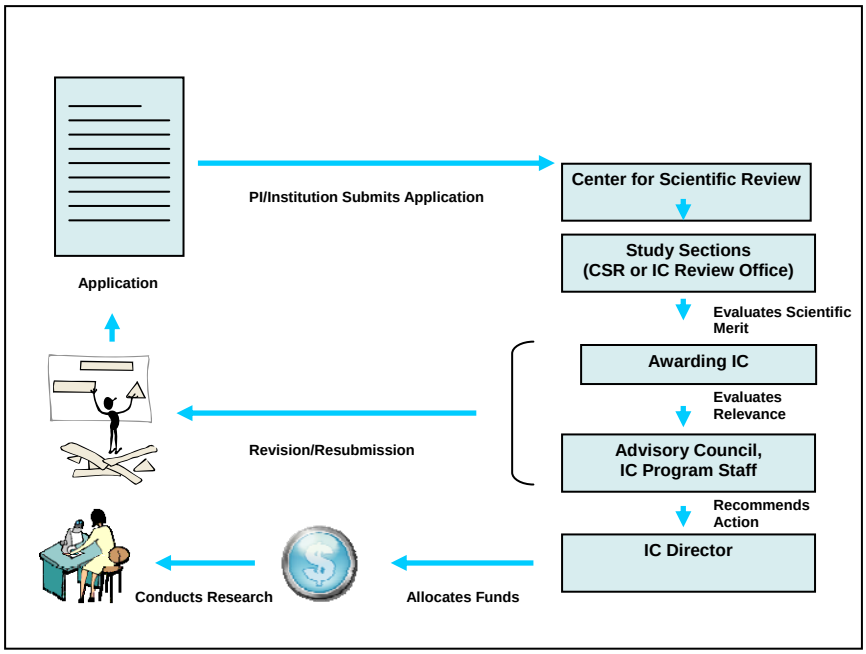


Figure 2: The two-stage NIH peer review process

Within this general NIH funding-policy rubric, individual ICs have defined their own practices to meet the needs of the diverse science and health communities relevant to the NIH mission. All ICs receive and review the results from the first phase of peer review, but additional factors are also considered.

Among these are whether the applicant is an early-career investigator, the level of other support available to the investigator and potential scientific overlap, as well as scientific and public health needs and balance.

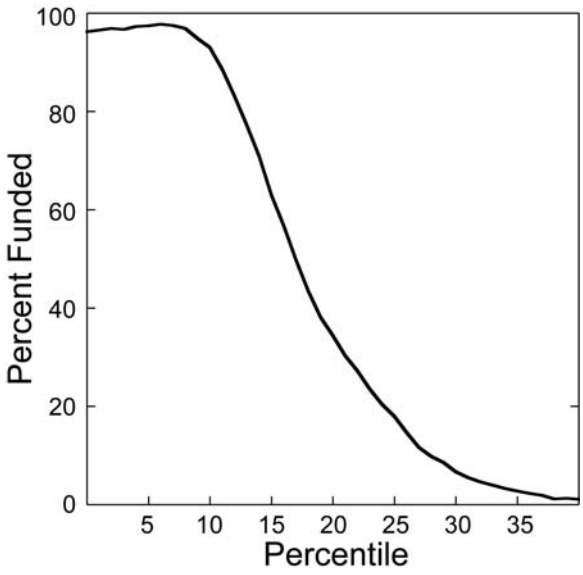


Figure 3. Relationship between NIH funding to percentile scores for R01 applications in fiscal year 2007. *Source: Dr. Jeremy Berg*

As a result, the NIH funding curve for any given year does not reflect an absolute drop-off based on percentile score (see Figure 3 for FY 2007 data). Rather, the breadth of this

curve is caused by differences in funding policies (and effective paylines) for different ICs, as well as the inherent breadth in individual IC funding curves.

DIAGNOSTIC PHASE: METHODS AND DATA COLLECTION

For the 2007-2008 peer review self-study, NIH ICs submitted comments, statements, and brief reports of peer review experiments, including those that were not successful. This input was solicited specifically from the leadership of all NIH ICs, based upon questions used in a broader NIH peer review survey. The NIH also obtained and examined information about peer review practices used by other domestic[3] and international[4] agencies to evaluate alternatives to conducting reviews, staffing review panels, post-review activities, and grant mechanisms.

IC responses to this specific query addressed issues spanning the entire peer review process from solicitation of applications, receipt, referral, review, and funding of grants, at varying levels of detail. Many of the experiments that were reported for this study have, by the time of this report, already been incorporated in standard practice in some ICs. Also, what are considered experiments in some ICs may already be standard practice in others. Several experiments involved the development and application of sophisticated electronic enhancements that facilitate the process by rapid transmission of information to many recipients and providing immense power for data and knowledge management.

Various electronic and many non-electronic experiments have been done or are ongoing to find more efficient and effective ways to conduct parts of the review and funding process at the NIH. These include: Internet-assisted review, videoconferencing, shortening the review cycle, shortening application length, criterion-based applications, and providing enhanced reviewer orientation. The recruitment of an adequate number of high quality reviewers remains one of the major issues in maintaining the overall quality of the NIH system of peer review, and is an area of concern that was prominent in the IC comments.

The many IC comments, suggestions and experiments--and their diversity in scope--demonstrate that a considerable amount of effort is being dedicated toward this effort (for more information, see *Challenges, Goals, and Recommended Actions* and *Appendix I, Previous and Ongoing Peer Review Experiments*). However, ICs comments also pointed to the importance of approaching the testing and implementation of system improvements in a coordinated way, to evaluate the impact of changes, individually and together, on the complex peer review system. ICs noted that the impact of too much change, too quickly, on the people who carry out the work must also be considered.

[3] Burroughs Wellcome Fund, Department of Defense Congressionally Directed Medical Research Programs, Gates Foundation, Howard Hughes Medical Institute, National Science Foundation, Robert Wood Johnson Foundation, Susan G. Komen Breast Cancer Foundation

[4] Australian National Health and Medical Research Council, Canadian Institutes of Health Research, European Research Council, Institut national de la santé et de la recherché medicale, Deutsche Forschungsgemeinschaft, Japan Society for the Promotion of Science, RIKEN, Singapore National Medical Research Council, Swedish Research Council, United Kingdom Medical Research Council

Center for Scientific Review: Ongoing Activities

Independent of the 2007-2008 NIH peer review self-study, CSR conducts its own ongoing analysis of feedback received directly from the broad scientific community regarding the peer review process. In response to this input, CSR has initiated its own peer review experiments and policy changes to address the following areas: i) improving study section alignment and performance; ii) shortening the review cycle; iii) recruiting and retaining more high-quality reviewers and decreasing the burden on applicants and reviewers; and iv) improving the identification of significant, innovative and high-impact research. Any overlap between the peer review self-study process and CSR's efforts will be noted within this document, where appropriate.

These efforts will continue along with the analyses of IC experiments (for more information, see *Appendix I, Previous and Ongoing Peer Review Experiments*). CSR is currently conducting pilot programs in the following areas:

- Asynchronous Electronic Discussion (AED) is just starting to be used and is in pilot testing in collaboration with some ICs.
- Videoconferencing is beginning to be used for small reviews.
- An automated referral system is being developed that conducts reviewer assignments by electronic procedures.
- Several pilots have been initiated to evaluate the efficacy of shorter applications.
- In early 2006, all summary statements for new R01 investigators were posted within 10 days of the study section meeting, and the summary statements for all investigators within 30 days of the study section meeting.
- Beginning in 2007, all summary statements were released on this expedited schedule.
- Beginning in 2006, senior CSR management reviews IRGs biannually to adjust study sections commensurate with changes in science.
- Bi-monthly open house workshops are also being held to solicit feedback from leaders of the scientific community and other stakeholders to determine i) how the current study section alignment serves science, and ii) how well newly emerging research areas can be served by peer review.

During the diagnostic phase of the 2007-2008 peer review self-study, the NIH canvassed its stakeholder communities, including the extramural community, patient advocacy groups, voluntary health organizations, professional societies, and NIH staff, to gather a broad set of ideas for enhancing the peer review system. This process included posting an online Request for Information (RFI), an NIH-internal survey, and an interactive Web site for liaisons; collecting data from previous and existing NIH peer review experiments and practices; direct communication with stakeholders through teleconferences, email, and letters; and hosting a series of internal and external consultation meetings and regional meetings across the nation. Many source documents containing information that has been gathered during the diagnostic phase of the NIH peer review self-study process are available online (http://enhancing-peer-review.nih.gov/).

Consultation Meetings

The consultation meetings followed a variety of formats to collect opinions and recommendations, and to foster discussion. Several meetings, for example, invited participants to present statements and proposals offering specific strategies or tactics for enhancing NIH peer review and research support. As appropriate during this iterative process, NIH staff presented emerging themes for consideration and discussion.

The consultation process revealed a remarkable amount of resonance, within and outside the NIH, on the key challenges facing the peer review system. Many key themes were repeated in independent meetings; these are thematically grouped and listed below. In addition, the consultations clarified that most peer review challenges are multi-factorial, and nonlinear in their capacity to respond. Thus, combining recommendations that emerge from the peer review self-study will likely be the only way to effect significant and lasting change.

Extramural Research Community Consultation

In fall of 2007, the NIH convened three regional meetings to gather input from the extramural community. Prior to these meetings, attendees were invited to prepare and present brief presentations to offer specific strategies or tactics for enhancing NIH peer review and research support. Full summaries of these meetings are posted online (http://enhancing-peer-review.nih.gov/). Key themes that emerged from these discussions are presented below as they relate to three general categories: core values/review quality, criteria/scoring, and career stages.

Core Values/Review Quality

- Minimize review bias by providing reviewer training and evaluating reviewers.
- Incentives, such as increased grant support or service flexibility, may enhance the recruitment of qualified reviewers.
- Matching expertise to proposal content is critical to review quality.
- Increased flexibility in study section constituency, electronic tools, and/or external input from professional societies may help obtain appropriate review panel expertise.
- Proposal resubmissions should be re-reviewed by the original review panel, not a new study section.
- The group dynamic of face-to face-review meetings broadens review and encourages honest discussion.
- Two-tier review may enable a more thorough and potentially less biased, evaluation of a proposal, but may increase administrative burden.
- Adequate diversity, and in some cases, public review participation, provides unique and necessary perspective to the review process.
- "Blinding" applicants, if truly achievable, may reduce review bias.

- Limiting the number of grants per NIH-funded investigator, and potentially increasing flexibility in their size and duration, would help streamline review and unclog the system.
- Setting salary limits on NIH grants would have significant impact on the peer review process.
- Shortening application length should reduce administrative burden and enhance the overall quality of review.
- Shortening application length may disadvantage some investigators, such as first-time applicants, or certain areas of research, such as clinical research.
- Enhanced recognition of true innovation, albeit difficult to define objectively, would reduce the current peer review conservatism.

Criteria and Scoring

- Minimize review bias by standardizing review criteria/scoring.
- Eliminate unscoring; all applicants deserve clear and unambiguous feedback on grant submissions.
- Multiple criteria and/or scores would provide increased dimension to review.
- Dominant personalities or poor study section leadership can bias review discussions and outcomes.
- Letters of intent, pre-proposals, and/or opportunities for applicants to correct factual errors may streamline the review process.
- Ranking proposals at the end of a review session would help reduce bias by calibrating and normalizing the application scoring process.

Career Stages

- Only seasoned, established investigators should staff review panels.
- Early-career investigator applications should be reviewed separately.
- Retrospective review may be appropriate for established investigators.

Advocacy Group Consultation

On October 22, 2007, the NIH convened representatives of advocacy groups to discuss issues related to peer review. Prior to the meeting, attendees were invited to prepare and present brief presentations to offer specific strategies or tactics for enhancing NIH peer review and research support. A full summary of the meeting is available online (http://enhancing-peer-review.nih.gov/). Key themes that emerged from attendee presentations and the ensuing discussions are presented below.

Core Values/Review Quality

- Minimize review bias by providing reviewer training and evaluating reviewers
- Matching expertise to proposal content is critical to review quality.

- Increased flexibility in study section constituency, electronic tools, and/or external input from professional societies may help obtain appropriate review panel expertise.
- Two or more tiers of review may enable a more thorough and potentially less biased, evaluation of a proposal, but may increase administrative burden.
- Adequate diversity, and in some cases, public review participation, provides unique and necessary perspective to the review process.
- Enhance the NIH's role in the funding process by linking review to the NIH mission.
- Limiting the number of grants per NIH-funded investigator, and potentially increasing flexibility in their size and duration, would help streamline review and unclog the system.
- Shortening application length should reduce administrative burden and enhance the overall quality of review.
- Shortening applications may disadvantage some investigators, such as first-time applicants, or certain areas of research, such as clinical research.
- The review of clinical and basic research should be separate.
- The review process is too conservative and does not encourage enough risk-taking.
- Milestones and deliverables could establish structure to the review process, but may encourage conservatism.

Criteria and Scoring

- Minimize review bias by standardizing review criteria/scoring.
- Review criteria should include impact, public health relevance, and translational relevance.

Career Stages

- Only seasoned, established investigators should staff review panels.
- Early-career investigator applications should be reviewed separately.

Professional Societies Consultation

On July 30, 2007, the NIH convened representatives of professional societies to discuss issues related to peer review. Prior to the meeting, attendees were invited to prepare and present brief presentations to offer specific strategies or tactics for enhancing NIH peer review and research support. A full summary of the meeting is available online (http://enhancing-peer-review.nih.gov/). Key themes that emerged from these discussions are presented below as they relate to three general categories: core values/review quality, criteria/scoring, and career stages.

Core Values/Review Quality

- Minimize review bias by providing reviewer training and evaluating reviewers.

- Incentives, such as increased grant support or service flexibility, may enhance the recruitment of qualified reviewers.
- Matching expertise to proposal content is critical to review quality.
- Interdisciplinary research and some career stages have unique review needs.
- The group dynamic of face-to face-review meetings broadens review and encourages honest discussion.
- Two-tier review (such as editorial board review) may enable more thorough and potentially less biased review but may increase administrative burden.
- The huge number of NIH grant mechanisms is confusing and leads to unnecessary complexity in the review process.
- Shortening the review cycle, and potentially application length, would mitigate information overload.

Criteria and Scoring

- Minimize review bias by standardizing review criteria/scoring.
- Eliminate unscoring; all applicants deserve clear and unambiguous feedback on grant submissions.
- Dominant personalities or poor study section leadership can bias review discussions and outcomes.

Career Stages

- Only seasoned, established investigators should staff review panels.
- Interdisciplinary research and some career stages have unique review needs.

NIH Internal Consultation

The NIH hosted three consultation meetings during the summer and fall of 2007. These meetings invited NIH staff to convene and discuss issues related to peer review. Full summaries of the meetings are available online (http://enhancing-peer-review.nih.gov/). Key themes that emerged from these discussions are presented below as they relate to three general categories: core values/review quality, criteria/scoring, and career stage.

Core Values/Review Quality

- Minimize review bias by providing reviewer training and evaluating reviewers.
- Incentives, such as increased grant support or service flexibility, may enhance the recruitment of qualified reviewers.
- Better alignment of application structure and review criteria would enhance review quality.
- Improved communication between applicants, reviewers, and NIH staff would enhance review.
- Interdisciplinary research and some career stages have unique review needs.

- Identifying innovation is important but difficult due to its variable definition and interpretation among applicants and reviewers.
- Practice and opinion vary across the NIH regarding the respective roles of review and program staff in the overall peer review process.
- Evaluation of existing NIH peer review practices requires defining and articulating optimal review outcomes as well as impact on applicants, reviewers, and NIH staff.
- Psychometric analyses will help account for the human element in the practice of peer review.

Criteria and Scoring

- Minimize review bias by standardizing review criteria/scoring.
- Eliminate unscoring; all applicants deserve clear and unambiguous feedback on grant submissions.
- Multiple criteria and/or scores would provide increased dimension to review.

Career Stage

- Interdisciplinary research and some career stages have unique review needs.

Formal Analysis of Input from RFI and Other Correspondence

The NIH enlisted a contractor to collect, organize, and analyze responses from the research community and the public to the request for ideas and recommendations on how to enhance the peer review process. The final report resulting from this effort describes methodology used to extract meaning from the text of 2,803 submitted comments from the RFI (posted online at http://enhancing-peer-review.nih.gov/), the NIH-internal survey, and email, telephone, and letter correspondence (17). Each narrative response varied in length from several lines of text to 10 pages of single-spaced text, with the average length being about 2.5 pages of single-spaced text. A hermeneutic approach was used to derive from these comments a "peer review" of the peer review process itself. This overall process thus transformed an initial narrative from RFI respondents into an organized list of new ideas and suggestions about peer review. A total of 2,724 records were received after duplicates were identified and marked as such in a project database. Analyzing a sample of the responses enabled a coding scheme to be iteratively generated. This analysis revealed the top issues by number of coded meaning fragments (Table 1[5]). Note that the top ten peer review comments ("quotes" in Table 1) are the same for both analyses with minimal changes in the rank order (with the exception of "Funding").

Table 1. Comparison of rank order for comments analyzed at 20,000 quotes and 40,000 quotes.
Source: Ripple Effect Communications, Inc.

Peer Review Code Category	**1st Run – 20,000 Quotes**			**2nd Run – 40,000 Quotes[6]**		
	Rank	**Count**	**Percent**	**Rank**	**Count**	**Percent**
Reviewers	1	3,118	19 %	2	5,097	19 %
Application Process + Format	2	2,798	17 %	1	5,908	
Score	3	2,419	15 %	4	2,876	22 %
Selection	4	1,623	10 %	6	1,623	11 %
People in Review Process-Investigators	5	1,272	8 %	5	2,439	6 %
Careers-New Investigators	6	1,193	7 %	7	1,597	9 %
Funding-Number of Grants + NIH Too Little Funds	7	1,172	7 %	3	3,823	6 %
Review Staff	8	820	5 %	10	820	14 %
Award Mechanisms	9	812	5 %	8	1,490	3 %
Criteria	10	811	5 %	9	941	6 %
Total		16,038	100 %		26,614	100 %

[5] See http://enhancing-peer-review.nih.gov/meetings/Peer_Review_Report_2007_12_03v3.pdf for full report.
[6] To verify the consistency in coding, the first 20,000 records were analyzed and then compared to the analysis of the larger set of 40,000 records.

This effort revealed that issues relating to reviewers were the top priority among respondents. Regarding reviewers, the main concern was an insufficient number of quality reviewers to adequately review grant proposals, caused primarily by a lack of incentives for quality reviewers to take valuable time away from doing their own research. Public comments noted that the application process was too time-consuming, and that electronic means could be used more often in the review process to minimize travel time and expense. Lack of useful feedback to investigators was cited as a related problem. Deficiencies noted in the scoring process included inconsistency, based primarily on lack of uniformly adopted standards of review, reviewer bias, and inaccurate calibration of review scores with assessing scientific quality and making funding decisions. Statistical analysis of the data was also conducted, resulting in an organized list of ideas and suggestions for peer review adjustments and potential pilot programs (17).

CONSIDERATION OF INPUT: THE COMMUNITY VIEW

The two NIH-convened peer review working groups, the Advisory Committee to the NIH Director Working Group on NIH Peer Review (ACD WG) and the Steering Committee *Ad Hoc* Working Group on NIH Peer Review (SC WG), each met independently several times between November 2007 and February 2008 to evaluate and discuss input received during the diagnostic phase of the peer review self-study. Discussion at these meetings considered all information obtained from NIH staff and stakeholders to define the major challenges, articulate potential solutions, and prioritize selected actions likely to have the most transformative effects on the NIH peer review system.

Themes that emerged during the consideration of input from the peer review self-study diagnostic phase are presented below in context of the discussion surrounding these issues.

The NIH must assure the core values of peer review.

Any proposed actions to enhance peer review must maintain scientific competence, fairness, and integrity—thus preserving the core values of peer review. It is also critical that proposed solutions address a specific problem, do not introduce additional administrative burden, and be testable in some way. Many cautioned that NIH should carefully consider the impact and potentially unintended effects of peer review interventions on applicants, the scientific community, and NIH staff.

Overview of potential interventions

Several potential key actions were considered that met the aforementioned criteria. These include: i) rating applications in a scientifically defensible manner; ii) standardizing the criteria for evaluating proposals; iii) shortening application length; iv) limiting minimum investigator percent effort per grant; v) decoupling "guidance" from reviewers by restricting direct responses of applicants and reviewers to any previous reviews; vi) tailoring review to career stage; and vii) creating and maintaining an explicit pathway for funding innovation or uniqueness.

During discussions with the community, several models for review implementation were proposed and discussed. One class of these, variations of editorial board-like models of peer review, met with considerable enthusiasm. While it may not fit all types of grants, many see this approach as a way to identify originality, increase the breadth of review, and reduce emphasis on minutiae. One key concern is scalability and potential burden on NIH staff. Another worry is that the editorial board model may contribute to scientific elitism; this may be countered to some degree by careful selection of a diverse set of reviewers per study section. In all cases, defining the "best" science is quite context-dependent.

The community expressed enthusiasm for setting limits (on numbers of grants, duration of grants, or minimum percent effort) as another strategy to reduce the total number of

applications in the system. Focusing on track record may elevate review discussion by underemphasizing method and detail, potentially alleviating review burden.

The cornerstone to review quality is recruiting and retaining a cadre of excellent reviewers.

Improving review quality must first address the problem of enhancing the culture of review: Several actions were considered: i) improving the review system so that those who serve feel that their time is well spent and they are contributing effectively; ii) instilling the expectation that those who are invited to serve will do so; and iii) considering appropriate incentives for service that are attractive yet do not make study-section service an entitlement. Reviewer training could elevate review discussion discourse and help to level the consistency among study sections. There is also general agreement that the best reviewers are relatively senior scientists well into their careers. While requiring reviewers to have NIH funding at the time of service may enhance the culture of review, such a requirement may constrain the size of the reviewer pool. Providing special incentives to study-section chairs may enhance recruitment, although some are concerned that this could be costly and potentially create entitlements. To ensure accountability, the NIH should have a system in place to review reviewers.

Ranking can reduce bias, especially if used as part of a two-stage, or multiple-criteria review.

One of the most important actions toward enhancing peer review is developing a rating system that is fair and consistent. In addition, there is a need to derive as much information as possible from a review. Various possible methods have been introduced to improve review quality and also reduce bias introduced by reviewer style heterogeneity. These include ranking, binning, and weighting criteria. Increasing the ability to explicitly measure and reward impact may be a transformative step forward.

Ranking was discussed in several contexts. While there is great support for a "ranking only" approach, the need to standardize rating across multiple study sections diminishes the feasibility of this option. Moreover, ranking may be applied before or during a review meeting. Thus, employing a range of schemes for explicitly ranking applications has the potential to enhance recognition of potential impact and to minimize unevenness during the course of review meetings as well as the influence of individual reviewers.

The use of multiple-criteria scoring, to explicitly measure multiple dimensions of an application, could help ICs to make funding decisions using the information provided during review. Potential criteria include impact, innovation/originality, the track record and other qualities of the investigator(s), approach, and the characteristics of the research environment. Importantly, the application structure must match the review criteria to minimize administrative burden and ensure review quality. Providing and/or weighting multiple criteria would not preclude determining an overall application score. Combining ranking with other potentially transformative actions, such as shorter applications and/or two-stage board review, may have maximum impact on enhancing peer review.

Every applicant should receive a clear assessment of the scientific merit of his or her proposal.

Providing feedback to all applicants is an important part of ensuring review quality and fairness. However, there is strong consensus that it is not the job of a reviewer to guide applicants by suggesting methodological fixes to a proposal that is inherently flawed and non-competitive with respect to potential impact or other key issues.

For a relatively small subset of applications, employing a "Not Recommended for Resubmission," or NRR, label would send a clear message to applicants and could help reduce the currently clogged review system. Implementing this concept could contribute to decreasing administrative burden.

Applications from early-career investigators must be funded at an appropriate level.

There is general consensus within the community that early-career investigators should be compared with each other in review. However, consensus does not exist on whether a targeted award is better than ICs funding a set threshold of early-career investigators, or whether early-career investigators should receive special consideration at all. Some view special programs as doomed to fail, in that they stigmatize the recipients, potentially creating funding problems later. Those holding this view recommend instead providing meritorious early-career investigators with independence--the freedom to choose and conduct research with few other demands--as early in their careers as possible. Others argue that unless a different type of review is provided--one that encourages risk-taking, early-career investigators quickly acculturate to "safe science," thus impeding their creativity and future scientific growth.

Workforce analyses to date suggest that interventions of some sort will be required to compensate for the aging of the researcher pool. Currently, OER has enlisted actuarial expertise and is conducting dynamic modeling exercises of the biomedical workforce. The outcome of these analyses will inform the NIH's actions related to support of scientists at all stages of the career pathway.

Diverse types of science are needed for progress toward improving human health.

Different types of science may deserve different modes of review. For example, clinical research and team science have very different components and needs than do investigator-initiated basic research projects done primarily in a single laboratory. There is also the problem of comparing dissimilar proposals. Clinical research that is labor- and cost-intensive, and long-term, is very different from relatively simple, inexpensive, and straightforward studies in model organisms. However, each has high value for biomedical research. Given these challenges, it is especially important to be sure that well-designed clinical studies get a fair review and a reasonable chance for funding if they fulfill the NIH mission and have high public health value.

Recognizing and rewarding true innovation/originality is a topic of great interest within and outside the NIH, generating substantial discussion in consultations with stakeholders. The term innovation is notably difficult to define with precision. Throughout this report, the term is intended to refer to research that cuts new ground, from a conceptual or technical perspective, or is strikingly distinct or even unique compared with other ongoing research. Some have suggested creating an explicit pathway for innovation and/or transformative research--the latter designed to create new paradigms. One approach would be to establish a "transformative" R01 grant whose application and review process would differ markedly from the current R01. This approach would in many ways resemble the current NIH Director's Pioneer Award program, with a brief, essay-based application that is person-focused and features a streamlined biographical sketch. Instituting any sort of innovative or transformative track would send a message to the scientific community that the NIH is serious about funding highly innovative work, and is willing to take risks and accept "failure" of many ideas. For this idea to work, the NIH must be very clear about its intent and adhere to its goals. Concern about spending NIH funds on risky projects could be addressed by monitoring such projects with benchmarks. The NIH should also be clear that other types of studies besides those which are "transformative," including incremental, data collection, and tool-building research, are not only important but necessary for progress.

Institutions should assume more responsibility for nurturing faculty and supporting the research enterprise.

There is general agreement within the community that institutions should remain committed to their faculty, without shifting substantial financial burdens such as salary to the NIH system. The heterogeneous reliance of institutions on grant money for investigator salary support contributes significantly to this problem. The potential impact of limiting NIH salary support to a defined percentage could be large. However, instituting any changes in this realm would be a fundamental shift in NIH philosophy, since the NIH values institutions that value research. It may have severe, unintended consequences since institutions are built on many different business models. Some investigators, for example, may be forced into additional administrative work at their institutions, which would reduce time for research.

Although this issue transcends peer review, many in the community feel strongly about making a statement that institutions receiving NIH funding should more uniformly accept financial responsibility for their faculty.

Peer review provides information about scientific merit, but the NIH makes funding decisions that are also influenced by portfolio balance and public health need.

The ambiguity and subjectivity of the term "best" science has created confusion about the role of study sections in the overall peer review process. While review provides information about scientific merit, the NIH makes funding decisions that are also influenced by portfolio balance and public health need. Ultimately, funding decisions are made by the NIH, and peer review enhancements that enrich the capability of the NIH to

obtain as much information as possible from a review are likely to be the most successful. It should be noted that the influence of the second stage of peer review, conducted by IC advisory councils, also varies among ICs, in part due to constituency pressures. The enduring goal is to balance scientific merit with the NIH mission, and ICs should be encouraged to use a range of tools to accomplish this.

Rating a proposal's responsiveness to criteria, potentially through weighting, is one approach that could increase the level of information available from a review. This could make it easier for the NIH to consider public health and scientific need when making funding decisions.

The most transformative actions to enhancing peer review will involve combining several interventions.

Working group discussions helped to cull the most significant challenges of enhancing the peer review process. The discussions also led to the formulation of various potential solutions to address concerns in these major areas. Many of these proposed solutions would have a synergistic effect when combined, and some may not work at all in isolation. This reality reflects the interdependency of review with other core processes in the biomedical research enterprise.

One action that is particularly integral to many other proposed ideas is shortening application length. Some ongoing efforts, such as the NIH Director's Pioneer Award program, have already introduced the concept of a much briefer application. Although the preliminary view is that this concept has met with enthusiasm and success, the Pioneer Award program is quite different in many aspects from typical R01 grants, and thus it may be difficult to extrapolate the impact of shorter application length to other funding mechanisms. In addition, while many researchers and NIH staff embrace shorter applications, some members of the community are very uneasy about what they view as a dramatic change. They express the concern that unintended and unforeseen consequences could compromise the core values of the peer review system. Perhaps one size does not fit all: Certain types of research, such as clinical studies and perhaps new uses of emerging technology, may not be adequately described in a relatively brief application. Using appendices or some similar approach is one way to address this issue. It may also be that a diversity of mechanisms is needed for different fields of science. In examining the issue of application length, it must be kept in mind that the use of shorter applications may enable a larger number of reviewers to read each application and, hence, to participate in review in a more informed manner.

CHALLENGES, GOALS, AND RECOMMENDED ACTIONS

The 2007-2008 NIH peer review self-study identified seven major challenges:

- Challenge 1: Reducing Administrative Burden on Applicants, Reviewers, and NIH Staff
- Challenge 2: Enhancing the Rating System
- Challenge 3: Enhancing Review and Reviewer Quality
- Challenge 4: Optimizing Support at Different Career Stages
- Challenge 5: Optimizing Support for Different Types of Science
- Challenge 6: Reducing Stress on the Support System of Science
- Challenge 7: Meeting the Need for Continuous Review of NIH Peer Review

In this section of the report, each Challenge is described and addressed with Goals and Recommended Actions. Challenges are designated with a numeral followed by a letter, to denote a sub-challenge (*e.g.*, Challenge 2A). Goals are not numbered or lettered, but each goal is followed by a main Recommended Action that has been placed in a box for emphasis. Specific recommended actions are listed below their parent recommendation and numbered sequentially (i, ii, iii).

For brief summaries of previous and ongoing peer review experiments at the NIH and other organizations, see Appendix I.

Challenge 1

Reducing Administrative Burden on Applicants, Reviewers, and NIH Staff

Despite recognizing the merit of the NIH peer review system, many view the process as overly burdensome. For many investigators, staying funded is a time- and labor-intensive exercise that can compromise the practice of research.

***Challenge 1A*: Too many applications in the system burden applicants, reviewers, and administrative staff.**

Beginning in 2002, the number of applications submitted to the NIH began to increase dramatically, transcending the historical growth rate (Figure 4). While it appeared that the numbers had reached a plateau of just under 80,000, CSR projects that this ceiling will be breached in FY 2008.

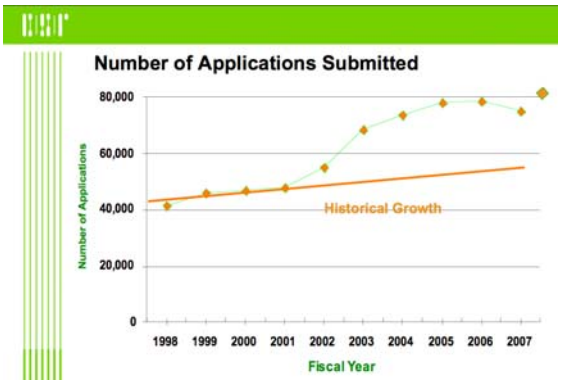


Figure 4. Number of NIH applications submitted compared to historical growth. *(Note that the number for FY08 is a CSR projection only). Source: CSR*

One effect of this application surge has been a steady increase in the number of *ad hoc* reviewers, rising to a high of approximately 15,000 in FY 2005 (Figure 5). Most recently these numbers have been reduced to approximately 12,000, due in part to a modest increase in the number of charter study section members. Many stakeholders noted the inherent drawbacks of using temporary reviewers who remain for only a short period of time, yet vote for all applications that are considered while they are in attendance.

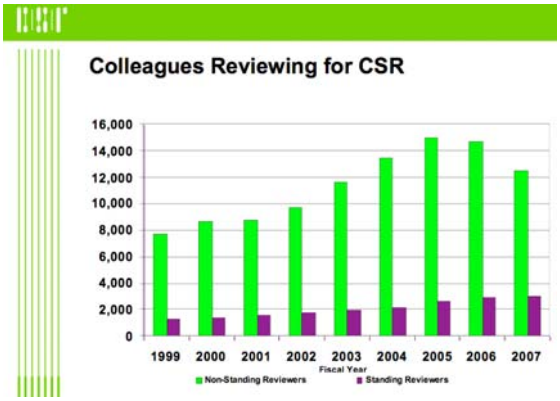


Figure 5. Comparison of the total number of chartered and *ad hoc* reviewers. *Note: Standing reviewers attend multiple study sections (not reflected in the chart). Therefore, per study section, the ratio of standing reviewers and non-standing reviewers is similar. Source: CSR*

Another trend contributing to the need for additional reviewers has been the attempt to reduce individual reviewer workload. The average number of applications reviewed per reviewer has steadily decreased from just slightly less than 12 applications per reviewer in the mid-1990s to a low of six applications per reviewer in 2005 (Figure 6). Over the past 2 years however, the trend appears to be increasing again: Each reviewer is now reviewing, on average, seven applications.

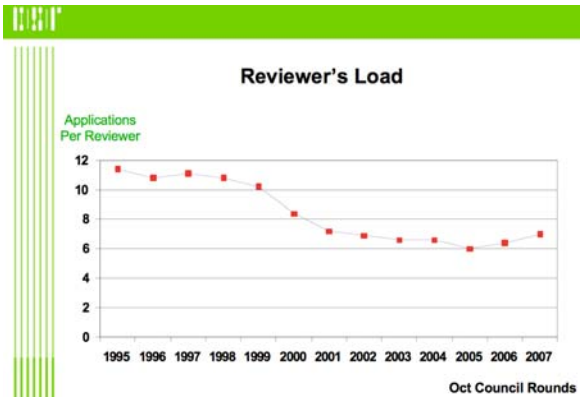


Figure 6. The number of applications per reviewer has decreased, requiring more total reviewers. *Source: CSR*

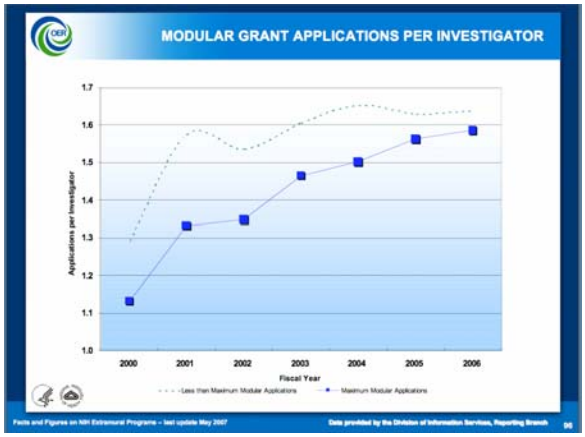


Figure 7. The number of modular applications per applicant has increased. *Source: OER, Division of Information Services*

Individual applicants increasingly apply for more than one application. In FY 2006, each investigator submitted, on average, 1.6 modular grant applications (Figure 7). Despite the multiple factors leading to this rise, this increase results in a greater burden to applicants, reviewers, and NIH staff.

<u>Goal</u>: To reduce the number of applications that need to be submitted by helping applicants make faster, more informed decisions to either refine an existing application or develop a new idea

Recommended Action: Provide unambiguous feedback to all applicants.

i) Establish a Not Recommended for Resubmission (NRR) decision option.

(see also Challenge 2, Enhancing the Rating System)

Summary statements present the strengths and weaknesses of each reviewed application. Reviewers may be asked to refrain from providing advice about a possible resubmission even when it is clear that an idea cannot be improved enough to make it competitive for payment. Currently, if in the course of review discussion an application is found to lack significant and substantial scientific merit, the study section can designate the proposal as "not recommended for further consideration (NRFC)." This action may also be recommended when serious hazards or unethical procedures are involved. No priority score rating is recorded, and the application's budget is not discussed, and the NRFC judgment results in an application's being ineligible for funding. In practice, however, the NRFC decision is rarely used.

The current rating system could be enhanced by the introduction of a stand-alone category checkbox entitled Not Recommended for Resubmission (NRR) for both scored and unscored proposals. The goal of this action would be to help applicants make faster, more informed decision whether to refine an existing application or to develop a new idea. A reviewer could check this category box in the event that he or she believes that a research idea would not have the appropriate potential impact or feasibility no matter how it was revised to be competitive in the future. Study-section consensus would be required, however, for an NRR designation to be applied to any given application. Establishing and implementing this category would intend to provide clearer, unambiguous information to an applicant; however, receiving an NRR decision would not prevent the applicant from submitting a revised, new proposal.

It is also recommended that the practice of unscoring be discontinued *(see Challenge 2, Enhancing the Rating System).* However, if the practice of unscoring of applications continues, there are expected to be circumstances where an unscored application would not receive an NRR decision; these would represent cases in which the proposal had merit but needed substantial modification to attain a potentially fundable score. Note that the NRR decision option is intended to provide clear feedback to applicants, whereas unscoring has been used as a mechanism for allocating review meeting time to focus on the most competitive applications.

ii) Provide ratings for all applications.

Particularly vexing to many is the current policy of not providing a score to a subset of applications that are viewed as not being competitive. First piloted in 1993, since 1995, all CSR regular study sections have employed the practice of unscoring. Most recently, the CSR average for unscored R01s has hovered around 50 percent. Yet, applicants (and their department chairs and deans) wish to have more information: specifically how close was the application to the unscoring cutoff point? Available data suggest that the majority of applications that were initially unscored fail to improve sufficiently in amended form to reach a fundable score (Table 2). A more recent analysis suggests that this trend continues. About 40 percent or less of unscored applications improve with subsequent amendments, and the average improvement has continued to decline over the past several years. The low rates of subsequent funding may also be due to reviewer bias from the previous reviewers' comments. However, reviewers, in general, successfully identify the

subset of applications that are not likely to improve significantly with amendment (Table 2).

It should be noted that for most unscored applications to improve, they would have to undergo marked improvement, whereas applications that were originally scored could undergo only modest improvement to be counted as "being improved." Although only a small percentage of unscored applications improve sufficiently as amended applications to reach a fundable score, given increasing pressures to secure funding and absent a clear message that an application is simply not competitive, it is not surprising that a significant fraction of investigators still submit amended applications. This effort could potentially be redirected toward other endeavors more valuable to the applicant.

Table 2. Comparison of funding rate for unscored and scored applications (*data not available).
Source: OER, Division of Information Services

FY	Number unscored/not initially funded	Unscored/funded on A1 (%)	Unscored/funded on A2 (%)	Number scored/not initially funded	Scored/funded on A1 (%)
1996	2,425	6.8	6.3	*	*
2000	2,898	9.0	7.6	*	*
2003	3,785	6.6	7.2	*	*
2005	4,791	3.7	7.2	6,756	31.0
2006	5,769	4.1	4.9	6,200	32.6
2007	5,076	2.6	2.9	4,840	18.8

In summary, the NIH should abandon the practice of unscoring applications. All proposals should receive a score, even those that are not fully discussed during a study section meeting, falling below whatever threshold has been established.

Challenge 1B: Increasingly, three submission rounds are necessary before an application is funded.

Initial (A0) submissions now fare poorly in peer review. Many members of the research community have reported the development of a review "queue." Given the finite nature of resources to fund applications, and the current system of allowing up to two amended applications per application, reviewers either conscientiously or sub-conscientiously have favored amended (A1 or A2) applications over A0 applications. The percent of R01 equivalent awards made to A0 applications has fallen from approximately 60 percent in FY 2002 to approximately 30 percent in FY 2007, with a corresponding increase in the percent of R01 equivalent awards made to amended applications (A1 and A2, Figure 8). Thus, increasingly, three submission rounds are necessary before an application is funded.

The deleterious consequences of this development are multifold. Support for meritorious science may be delayed if initial submissions are placed at the end of the queue. By considering amended applications, a reviewer may take an unintended stake in guiding an applicant: heeding a reviewer's advice may be rewarded with a better score. Conversely, if an applicant does not take a reviewer's advice, they may be penalized. This collective

behavior may lead to a score that is based on an applicant's responsiveness to reviewer guidance. It is interesting to note that in 1993, the Atwell Committee, charged with identifying potential ways to renovate the peer review system, asserted, "... summary statements should not be primarily tutorial in nature."(18). Another negative effect is that reviewers may unduly favor "last chance" (A2) applications, thus potentially leading to support of less meritorious science.

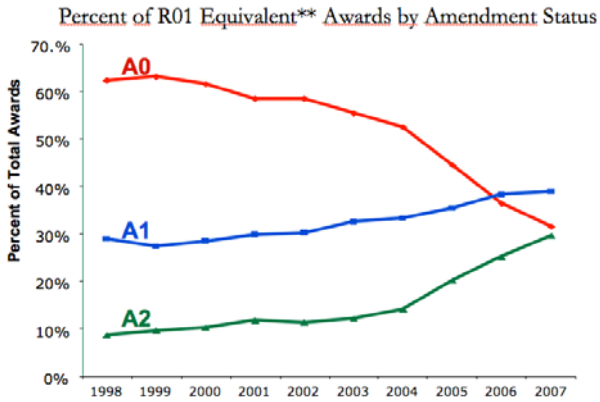


Figure 8. Inverse relationship between A0 and A1/A2 funding rate. *Source: OER, Division of Information Services*

***Goal*: To focus on the merit of the science presented in the application and not the potential improvements that may be realized following additional rounds of review**

Recommended Action: Eliminate the "special status" of amended applications.

i) Consider all applications as being new.

Study sections should focus on assessing the merit of proposed science as presented, as opposed to its potential after subsequent submissions. Considering all applications as new eliminates an applicant's expectation that following a reviewer's recommendations or "dictates" will improve his or her proposal's score. It also enables the formation of study sections that no longer need their standing members to be present at every round since all applications will be considered on the basis of their current scientific value. Thus, this practice would also remove the incentive to shift A0 applications to the end of the review "queue," refocusing evaluation on application merit rather than on ancillary factors, such as assigning "last-chance" status to A2 applications.

The potential consequences to eliminating the special status of amended applications are summarized below:

- The applicant:
  - may resubmit his or her application, with revisions as desired.
  - will no longer respond to reviewer comments as part of any subsequent submission. As a result, the application can be written more concisely.
- The reviewer:
  - will no longer see previous reviewers' comments.

  - will be able to write a more concise review.
- The NIH:
  - will no longer provide previous reviewers' comments.
  - will yield greater flexibility in study section makeup.

Note that both applicants and NIH will still distinguish between type 1 (new) and type 2 (competing continuation) applications.

There are numerous "one-time" applications already in use at the NIH (Requests for Applications, for example), and many foundations employ single, one-time only, solicitations. Coupled with unambiguous feedback, this approach could reduce the total number of applications that an applicant will need to submit.

ii) Identify the subset of applications for which corrective actions can be assessed without the need for re-review by a study section.

An alternate strategy would be if the corrective measures were judged to be appropriate, the application would not have to be placed back into the review queue, thereby markedly reducing burden on reviewers and applicants. To ensure compliance, the applicant's response would need to be incorporated into the terms and conditions of the award.

The National Institute on Deafness and Other Communication Disorders (NIDCD) employs an approach to administratively fund early-career investigator applications that require only modest revision. Applicants identified in this category are given the opportunity to write a 5-page response to present possible corrective measures, which is then evaluated by two IC council members. This practice, employed since 2001, has led to an increase in the funding of early-career investigators by NIDCD. For more information, see *Appendix I, Previous and Ongoing Peer Review Experiments*.

iii) Pilot the use of short, bidirectional "prebuttals" (for applicants and/or reviewers) to correct factual errors or explain factual items in review.

Factual errors in reviews are often challenging and time-consuming to address. Allowing review corrections prior to the first stage of review may alleviate some burden. Reviews could be posted on a secure electronic site. The applicant would be given a discrete window of opportunity to redress any factual errors in the posted review. The study section would have access to the prebuttal during the review process. Variants to this model include having reviewers direct specific queries to applicants before review meetings, or by requiring applicants to be on telephone standby during the panel meeting to answer questions. Reviewers could use this mechanism as a vehicle to ask for clarification or data on certain points they consider to be of key importance.

Note that the implementation of prebuttals would require adjustments in process and culture to allow sufficient time for prebuttal generation prior to review meetings.

(see also Challenge 3, Enhancing Review and Reviewer Quality)

Recommended Action: Shorten summary statements by focusing solely on the merit of science as presented.

Goal: To reduce application length to focus on impact and uniqueness, placing less emphasis on standard methodological details

Recommended Action: Shorten the length of the application and align it to specific review elements.

The NIH grant application is among the longest that is used by funding agencies and foundations worldwide. While the length provides applicants with great flexibility it also yields a daunting challenge to reviewers. Additionally many believe that providing too much flexibility has led to an overemphasis of being placed on fine detail, such as methodology. As a result, they argue, more important criteria such as the significance and/or impact of the proposed work becomes diluted. An NIH-issued RFI evaluated the community's views on this matter (NIH RFI NOT-OD-07-014). Of over 5,000 individual responses, 43 percent preferred a reduction to 15 pages, and an additional 27 percent opted for still shorter applications. Thus, in aggregate, 70 percent of all respondents preferred that applications be shorter than 25 pages (Table 3).

Table 3. Responses to NIH RFI NOT-OD-07-014. *Source: CSR*

Page-Limit Preference	Total	NIH Staff
5	5 %	1 %
10	22 %	13 %
15	43 %	41 %
25	27 %	45 %
No response	3 %	0
Total N	5,078	226

The majority of the respondents to this RFI did not believe that a shorter application would compromise their ability to present scientific ideas (68 percent responded that ideas could be communicated equally well comparing 25 to 15 pages, and 19 percent indicated that shorter applications would enhance communication).

A more detailed, sub-analysis (on 500 responses to this RFI) demonstrated that 49 percent of the respondents indicated that having to review an "equivalent number of shorter applications" would increase their willingness to review; however, they were not asked if they would review more applications. The majority of these respondents (65 percent) indicated that if the application were shortened, review criteria should be changed to emphasize ideas and/or impact. A third of those responding (in the sub-analysis) reported that shortening the application would not affect their ability to judge

the scientific merit of an application, and one-fourth thought a briefer application would actually enhance a reviewer's ability to judge scientific merit. In contrast, 32 percent felt it would be more difficult to judge the scientific merit of shortened application. Respondents also indicated that a shortened application would take either the same (27 percent) or less (50 percent) time to prepare. Thus, for at least half of those responding, shorter applications would translate to less burden.

Multiple ICs have piloted shortened applications. The National Heart, Lung, and Blood Institute (NHLBI), the National Institute of Allergy and Infectious Diseases (NIAID), and the NIH Director's Pioneer and New Innovator Awards, under the direction of the National Institute of General Medical Sciences (NIGMS), have used applications containing fewer than 25 pages. The use of shorter applications in these ICs seems to be favorable. For more information, see *Appendix I, Previous and Ongoing Peer Review Experiments*).

While the diagnostic phase of the peer review self-study did not engage in a detailed analysis of the optimal length of an NIH application, the overwhelming opinion (but not a unanimous view) was that a significant reduction is necessary to accrue the enhancements desired. Almost all members of both working groups favored reduction to somewhere between 7 and 15 pages. It should be noted that a shortened application will not necessarily be "easier" for an applicant to write. However, by distilling the application to its most significant elements--the potential impact of the work, the investigator(s) prior accomplishment, and the originality or innovation of the proposed work--a great deal of benefit will be gained by the applicant. Further, a shortened application will make it feasible for more reviewers to engage in review, and this should also be viewed as a very favorable benefit to the applicant. Importantly, the size of applications should be scaled to the complexity of the type of application (either mechanism or type of science[7]). Further details will be addressed during the implementation phase of the peer review self-study.

Challenge 1C: The proliferation of NIH funding mechanisms can be confusing and burdensome to applicants, institutional officials, and review staff.

The desire to meet the unique needs of different scientific communities has often led the NIH to create new mechanisms or for ICs to use mechanisms in slightly different ways across the NIH. While this flexibility is a potential benefit to investigators, administrators from applicant organizations report that this practice is confusing. In addition, many study sections review applications that will be sent to different ICs for funding consideration. As a result, reviewers are faced with the impossible task of reviewing applications that are being evaluated by ICs with varying goals and criteria.

[7] Clinical trial applications may have unique length needs, and appendices may be merited for these applications.

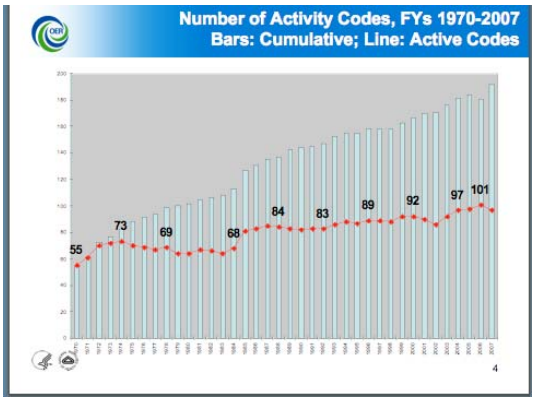


Figure 9. Number of activity codes from 1970 to 2007. *Source: OER, Division of Information Services*

Recommended Action: Where feasible, refine and harmonize existing mechanisms unless data-based evaluation suggests otherwise.

Challenge 2

Enhancing the Rating System

Unequivocally, the rating system that informs NIH peer review is central to every activity, and thus it is critical that the NIH consider carefully ways to ensure that scoring is both as accurate and informational as possible. Evaluating potential changes to the rating system may be particularly amenable to pilot testing.

Challenge 2A: Improve the usefulness of the rating system to inform decision making for both applicants and the NIH.

The most reliable, consistent rating system is one that reflects reviewers' abilities to discriminate. In current practice at CSR, each scored grant application is assigned a single, global score that reflects the consideration of five review criteria (significance, approach, innovation, investigator, and environment). The emphasis on each criterion varies from one application to another, depending on the nature of the application and its relative strengths. Individual reviewers mark scores (1.0 to 5.0, 41-point scale) to two significant figures (*e.g.,* 2.2), and the individual scores are averaged and then multiplied by 100 to yield a single overall score for each scored application (*e.g.*, 253). The best possible priority score is 100 and the worst is 500.

If in the course of review discussion, an application is found to lack significant and substantial scientific merit, it may be "not recommended for further consideration (NRFC)." This action may also be recommended when serious hazards or unethical procedures are involved. No priority score rating is recorded, and the application's budget is not discussed. The NRFC judgment results in an application's being ineligible for funding. In practice, however, the NRFC decision is rarely used.

Most research grant applications are also assigned a percentile rank. The conversion of priority scores to percentile rankings is based on scores assigned to applications reviewed during the current plus past two review rounds for standing committees. Percentiling ranks applications relative to others scored by the same study section, as an attempt to normalize scores between study sections. The NIH includes unscored applications in percentile calculations, so since the number of unscored applications varies by study section, including them affects the percentile distribution and attempts to make percentiling fair across study sections.

Rater reliability drops or fails to increase with a rating scale extended beyond 7 points (19). The NIH's current scale of 41 points (*i.e.*, 1.0 to 5.0) for initial scoring far exceeds that recommended by psychometric analysis. Further, once the scores are averaged and multiplied by 100, the resulting priority score appears to have more precision than it actually has. Less straightforward would be to use the current process, but round the score up or down to the nearest 10th. For example, a score of 157 would be rounded to 160. The current scoring system and process could be maintained and also accurately reflect the real number of significant digits by instituting the practice of rounding all

final, averaged scores to the nearest 10. For example, a score of 157 would be rounded to 160. Adopting such a system may lead to situations in which many applications near the payline have identical scores. This will result in greater dependence on program staff and priorities for funding decisions within similarly ranked applications. Similar potential actions include changing the scale initially used by reviewers to rate applications. A scale of whole numbers may be used (*i.e.*, 1 to 5 or 0 to 7), or priority scores could be determined by averaging reviewer scores without multiplying by 100. Any of these potential changes will lead to many applications near the payline having identical scores. Thus, NIH program staff will need to consider portfolio balance and IC mission for making funding decisions among similarly ranked applications.

Changes relating to the calculation, standardization, and reporting of scores are supported by a 2003 study, which also made recommendations regarding the NIH scoring system following the 1994 Government Accounting Office report that criticized some features of the review process at the NSF and the NIH (20). This report highly recommended using a procedure to standardize scores between reviewers by using z-scores: in essence, normalizing for reviewers that tend to score very high or very low. The study also suggested trimming down the point scale, and using disaggregated ratings.

***Goal*: To focus and elevate the level of discourse during study-section meetings**

***Goal*: To enhance consistency of rating and to engage all charter review members in the review of each application**

Recommended Action: Modify the rating system.

i) Rate multiple explicit criteria individually:

- Impact
- Investigator(s)
- Innovation/Originality
- Plan
- Environment (including information on institutional support for the applicant)

Criterion-specific rating provides flexibility for ICs to weight those criteria that are important to the mission and/or portfolio of the funding IC. Potential rating areas include impact, investigator(s), innovation/uniqueness, plan, and environment (including institutional support), as well as an overall score that is based on each reviewers overall sense of the importance of the proposal (see below). Explicit scoring criteria should be matched to specific sections of the application to directly address each topic. Note that the proposed criteria relate to R01 and similar research project grant applications and other variations may be more appropriate for other classes of applications. The "grain-size" to be employed and the level of discrimination desired will need to be developed during the implementation phase of the peer review self-study.

ii) Provide an independent, overall score.

Creating multiple ratings for explicitly stated criteria and then deriving an independent, overall score (that is not guided by an algorithm) may improve reviewer ratings in many substantive areas and would provide more detailed information, by criterion, on application quality. Some have argued that it would valuable to pilot a comparison between an algorithmic derived score versus an independent overall score; one key concern is that study sections may adapt to any formulaic approach by "gaming" the system.

Multiple ratings have been used effectively in the NIH Director's Pioneer Award program. Pioneer Award applications are rated by three criteria and an overall rating (Figure 10, left). Reviewers provide the overall rating, and it is not derived from the other criteria-based ratings. The four independent ratings provide information that can aid NIH program staff in discriminating within a population (Figure 10, right).

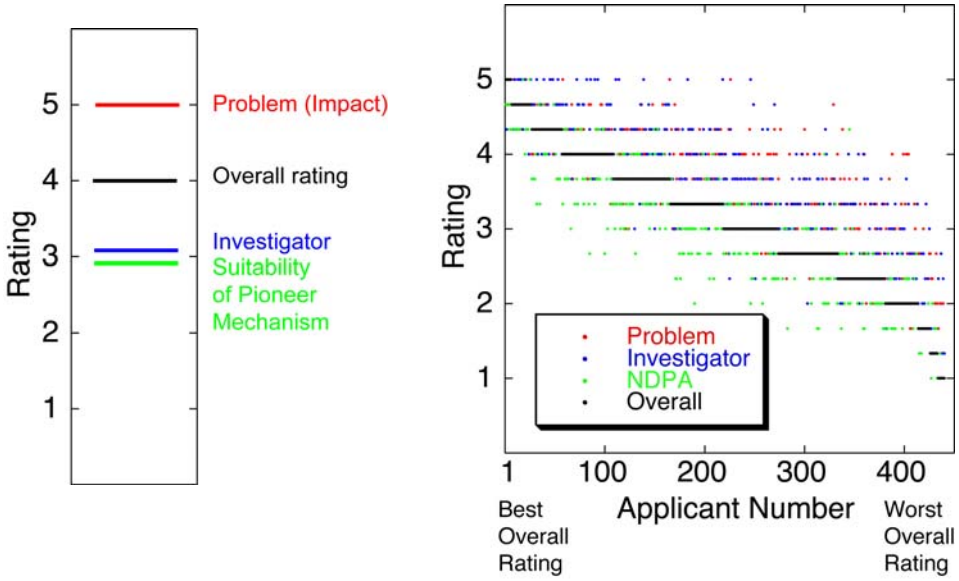


Figure 10. NIH Director's Pioneer Award rating scheme. *Source: NIGMS*

iii) Rank applications considered by the study section.

While there was almost unanimous consensus that a ranking system would offer great advantage to the current rating system, there was considerable diversity of opinion about when best to employ this tool. In one approach, reviewers would offer their ranking prior to the study section meeting. In one particular model espoused by Dr. Ken Dill of the University of California, San Francisco, reviewers would rank the applications assigned to them (this could be all the applications depending on the nature of the panel) and assign points based on where the application was ranked. This approach has the advantage of being consensus-independent, where one persuasive member does not unduly influence the deliberations. However, this approach would be difficult to normalize across study sections.

In a second model, charter review members would rank applications explicitly at the conclusion of the review meeting. Final rankings would be derived from frequency counts rather than averaging, and votes would be collected from everyone to prevent any single person's bias affecting the outcome. Analysis of the outcomes of testing this approach will provide a better understanding of what criteria are being considered by reviewers in arriving their final scores. The experience of other agencies and some foundations is that keeping study section members engaged throughout, and until the end, of the review meeting enhances the review process. Note that this approach will necessitate changes in review meeting culture, including the tendency for some reviewers to leave before the end of the meeting. Nonetheless, such changes have potential for substantial positive benefit.

The NIH Director's Pioneer (2004-present) and NIH Director's New Innovator (2007-present) Awards have used an alternate rating system that includes the explicit ranking of reviewers' top four applications. Reviewers have been generally pleased with the system. For more information, see *Appendix I, Previous and Ongoing Peer Review Experiments*.

***Goal*: To provide unambiguous feedback to applicants**

Recommended Action: Establish an unambiguous rating category: Not Recommended for Resubmission (NRR)

(see also Challenge 1, Reducing Administrative Burden on Applicants, Reviewers, and NIH Staff)

Recommended Action: Restructure the application to reflect the rating criteria.

(see also Challenge 3, Enhancing Review and Review Quality)

Measuring responses to several criteria is fairer to applicants and provides the NIH more information than is currently available from a study section meeting. This enriched information enables the NIH to consider public health and scientific need when making funding decisions, which helps the agency carefully steward its resources. Establishing and assessing multiple criteria may also increase review quality by helping a reviewer focus on diverse aspects of an application. For this to work, it is essential that the application structure match the review criteria to minimize administrative burden and ensure review quality.

The current application structure does not provide a clear location for an applicant to explain how his or her project meets each review criterion. Without a designated location for explaining impact, the prior accomplishments of the investigator, or innovation, applicants may fail to articulate important insights within the application. This omission also presents a considerable burden to reviewers who must search for explicit statements regarding impact, innovation, and other criteria. Moreover, the current review criteria may encourage a reviewer to spend time on the aspects of an application that are not critical to identifying the most meritorious proposals. Reviewers might also find panel

meetings to be a more positive experience if the focus was shifted to identifying strengths of applications, rather than weaknesses.

Challenge 3

Enhancing Review and Reviewer Quality

The cornerstone to review quality is recruiting and retaining a cadre of excellent reviewers. Concerns have been expressed by the broad scientific community that the practice of reviewing grants has lost its appeal, in part due to the evolving of science and the NIH budget flattening.

Challenge 3A: Improving review quality means addressing the larger problem of changing the culture of review.

Goal: To enhance review quality

Additional strategies can be employed to further enhance the review process and to yield reduced burden to reviewers, thereby facilitating their willingness to participate in a more meaningful way. Particularly appealing are two-stage review models in which the workload is divided, ensuring both requisite technical expertise and broad evaluation.

Recommended Action: Engage more reviewers per application.

i) Continue to pilot the use of two-stage review models.

Editorial board review models may particularly appropriate for large, complex applications. The basic premise of the editorial board review model—which mirrors journal review of scientific manuscripts--is that by engaging experienced reviewers to think broadly about the quality and impact of a proposal, the general discourse of review will improve. By enlisting outside experts for technical review as needed, less emphasis will be placed upon methods and detail, leaving more room for appreciation of the quality of the idea, and its impact on science and health. Two basic models have been proposed; they are not mutually exclusive and are presented here for the purpose of illustration only.

The first model employs a set of outside technical reviewers, who review all applications by mail. These technical reviews are distributed to board members and applicants prior to the study section meeting, to enable prebuttal, if necessary. This would allow applicants to correct factual review errors (only) and also offers a pivotal point for identifying proposals that are very clearly flawed from a technical standpoint. All board members would read all technical reviews, but only those designated as primary reviewers for a given proposal would write a concise, second-order "review of the reviews" for that proposal. These designated primary reviewers lead the review discussion, which focuses on impact and overall quality. This model would likely work well for large, multi- and interdisciplinary projects. However the numbers of individuals required make scalability to the whole of NIH very challenging.

The second model attempts to address potential scalability issues by specifying that only some applications (those requiring clarification/assessment of certain technical aspects) are sent out to external experts, for technical review. With this information (and perhaps prebuttal responses) in hand, the chartered board members would rank their assigned proposals. In a variant of this approach, as a first step, all charter review members would review every application for proposed impact. The subset of applications that lack sufficient impact could then be designated NRR.

A number of ICs have conducted peer review experiments that have either examined two-tier review models or employed pre-applications that resemble aspects of the proposed editorial board review model. For more information, see *Appendix I, Previous and Ongoing Peer Review Experiments*.

ii) Increase the use of electronic-assisted reviews.

For more information and several examples, see *Appendix I, Previous and Ongoing Peer Review Experiments*.

Recommended Action: Pilot the use of short, bidirectional "prebuttals" (for applicants and/or reviewers) to correct factual errors or explain factual items in review.

Factual errors in reviews are often challenging and time-consuming to address. Allowing review corrections prior to the first stage of review may alleviate some burden. Reviews could be posted on a secure electronic site. The applicant would be given a discrete window of opportunity to redress any factual errors in the posted review. The study section would have access to the prebuttal during the review process. Variants to this model include having reviewers direct specific queries to applicants before review meetings, or by requiring applicants to be on telephone standby during the panel meeting to answer questions. Reviewers could use this mechanism as a vehicle to ask for clarification or data on certain points they consider to be of key importance.

Note that the implementation of prebuttals would require adjustments in process and culture to allow sufficient time for prebuttal generation prior to review meetings.

(see also Challenge 1, Reducing Administrative Burden on Applicants, Reviewers, and NIH Staff)

***Challenge 3B*: Knowledge of the identity of an applicant/applicant's institution might bias reviewers.**

Consultations with the extramural community yielded the comment that the practice of "blinding" applicants, if truly achievable, may reduce review bias. The Publishing Research Consortium released an international study[8] on the perspective of peer review

[8] http://www.publishingresearch.net/PeerReview.htm

in scholarly journals (21). The majority of those surveyed (71 percent) had confidence in double-blind peer review, with 56 percent expressing a preference for this approach. However, several publications conclude that for journal articles, masking the identity of the authors had no impact on the quality of review (22-24). Moreover, there is a widespread view that the identities of the authors are difficult if not impossible to blind given a reviewer's knowledge of their field. However, it has been documented that after a policy change by the journal *Behavioral Ecology* to blind author identity, a significant increase in the number of female, first-author papers was observed (25). The applicability of this observation to peer review is supported by observations reported earlier regarding the peer review of postdoctoral fellowship applications in Sweden (26).

Recommended Action: Pilot anonymous review in the context of a two-level review system such as the editorial board model.

The NIH should consider the value of anonymous review; however, piloting this concept necessitates two-tiered (*e.g.*, editorial board style) review. The first, anonymous stage would assess scientific merit. The second, non-anonymous stage would take into account the investigator and his or her environment since these issues are critical to a project's ultimate success.

Challenge 3C: There is a need for standardizing reviewer, study-section chair, and scientific review officer training.

Reviewer training would elevate review discussion discourse and also help to contribute to review consistency among study sections. It would also help disseminate the vision of the NIH leadership. The broad scientific community expressed substantial agreement regarding key areas to be addressed in expanded training programs for reviewers:

- Emphasizing strengths, rather than weaknesses of applications
- Focusing on potential impact of the research rather than on standard methodology
- Reviewing the merit of the proposal and not re-writing it
- Recognizing the problem of implicit bias in study sections
- Using benchmark applications during panel meetings to provide review guidelines
- Point out potential bias towards lesser known applicant organizations

Several ICs have conducted experiments of reviewer orientation teleconferences to promote the consistent application of review criteria, particularly for complex applications. Favorable outcomes include: i) reviewers generally report improved ability to focus critiques on issues most germane to program announcements; ii) time spent discussing review and programmatic issues at review meetings is often reduced; iii) panel members report valuing the opportunity to discuss review and programmatic issues with NIH staff; and iv) conference calls often enhance interactions between NIH review and program staff. Several ICs reported that the benefits of orientation teleconferences outweighed the costs. For more information, see *Appendix I, Previous and Ongoing Peer Review Experiments.*

Recommended Action: Establish or enhance reviewer, study-section chair, and scientific review officer training.

i) Provide study-section chair training.

Given that study-section chairs are tasked with setting the tone and eliciting effective discussions during reviews, it has been proposed that they be provided with training in NIH core values, review criteria consistency, and meeting management strategies. Given the number of study-section chairs, it should be possible to provide these individuals with direct training. The cohort as a whole would benefit greatly from discussion of best practices among peers. It has also been proposed that incentives would attract the most qualified chairs although most expressed the view that the prestige of serving as a chair was sufficient for most people.

Reviewers would also benefit from standardized training. Given the number of reviewers engaged by the NIH, it has been suggested that the agency could create training materials that would be distributed to universities and research institutions or made available over a secure Web site. Individual reviewers could be asked to certify that they reviewed the training materials prior to appointment.

CSR currently provides extensive training for scientific review officers. Given the nature of the recommended actions, it is suggested that specialized training addressing those changes that are adopted be provided. This material should also be used by scientific review officers within ICs.

ii) Increase reviewer accountability.

Requiring some form of accountability for reviewers may improve the current system by increasing the quality of reviews and feedback given to applicants. It has been suggested that informal feedback to reviewers be provided, either by the chair and/or scientific review officer of the study section. An alternative approach would be through some group "rating" of each reviewer. Where egregious errors in process or judgment have been noted, more formal feedback should be required. The NIH should develop a standard process for providing this formal feedback, and in unusual circumstances, for the removal of reviewers from a study section.

Challenge 3D: There is a continued need to attract the most qualified ("best") reviewers.

During the 2007-2008 peer review self-study diagnostic phase, both the external and internal NIH scientific communities highlighted the challenge of recruiting high quality reviewers as a high priority area. While NIH grantees are not required to serve in a review capacity, all members of the scientific community should recognize that peer review service is crucial to the successful operation of the system. Although the percentage of associate and full professors serving as CSR reviewers remains at high levels (Figure 11), there is a perception that many of the most accomplished scientists do

not serve as reviewers. Contributing to the difficulty of recruiting quality reviewers from academia, industry, and other organizations is the perception that reviewer time is not being used effectively. Lowering logistical and cultural barriers may encourage all qualified reviewers to serve on study section panels. Various incentives may help to achieve this goal.

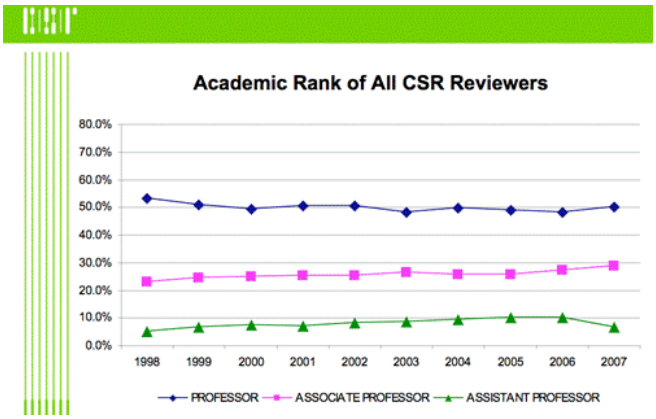


Figure 11. Academic rank of all chartered and non-chartered CSR reviewers. *Source: CSR*

The cohort of reviewers for the highly selective NIH Director's Pioneer Award are highly accomplished as measured by the percentage (127/174, or 73 percent) that have won awards in their fields or have received prestigious fellowships or honors (Figure 12).

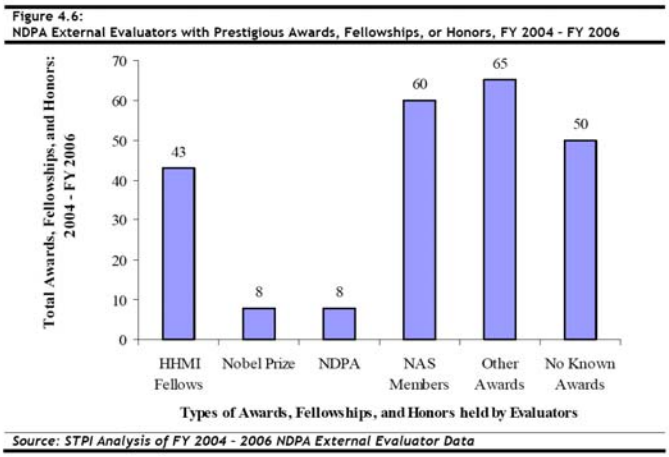


Figure 12. Awards and honors of NIH Director's Pioneer Award reviewers. *Source: NIGMS*

Therefore, it is possible to attract the most accomplished scientists to study-section service if the reviewers believe that the process is efficient and that their participation will have impact.

<u>Goal</u>: To enhance reviewer quality

Recommended Action: Create incentives for reviewers.

As part of the peer review self-study, the external scientific community recommended several means for potentially reducing reviewer burden and providing incentives to recruit high-quality reviewers. Many of these suggestions were also proposed by ICs, and most had been recommended to CSR and ICs prior to the peer review self-study process. As a result, several peer review experiments as well as procedural and policy changes have already been undertaken or have been implemented. For more information, see *Appendix I, Previous and Ongoing Peer Review Experiments*.

i) Allow more flexible review service.

Reviewer participation could conceivably increase with various types of incentives:

- Provide the opportunity for flexible review service (two meetings per year)
- Provide the opportunity to limit the term of service to 2 years
- Conduct study sections on a more regional basis around the country

ii) Provide more flexible grant-submission deadlines for all reviewers.

In 2008, CSR abolished submission deadlines for all study-section chairs. Providing greater flexibility in receipt dates for panel members may serve as an effective recruitment tool.

As part of the 2007-2008 peer review self-study, the external scientific community recommended several means for potentially reducing reviewer burden and providing incentives to recruit high-quality reviewers, such as allowing more flexible service (*e.g.*, twice per year) and providing reviewers more flexible deadlines for grant submission.

Recommended Action: Link potential study-section service to the most prestigious NIH awards.

(see also Challenge 4, Optimizing Support for Different Career Stages)

Adding a review-service condition to MERIT[9]/JAVITS[10]/NIH Director's Pioneer Award, and, potentially other, meritorious awards in the future, would recognize outstanding scientists with high impact in their fields and help to grow the pool of expert reviewers.

Challenge 3E: Is there adequate participation of clinician scientists in peer review?

Anecdotally, it is reported that clinicians, particularly those from surgical subspecialties, find it very difficult to serve on study sections, given their patient-care responsibilities.

[9] The principal feature of the MERIT (Method to Extend Research in Time) award is the opportunity to obtain up to 10 years of research support in two segments and thereby relieve awardees of the need to prepare frequent renewal applications.

[10] A Javits Neuroscience Investigator Award (R37) is a conditional, 7-year research grant given to scientists selected by the NINDS Council from among the pool of competing applicants during a given grants cycle.

The NIH needs to examine this issue to determine if there is a more optimal number of clinician scientists to participate in peer review.

Recommended Action: Ensure participation of adequate numbers of clinician scientists by analyzing patterns of participation by clinician scientists and by providing more flexible options for review service.

(see also Challenge 5, Optimizing Support for Different Types of Science)

Challenge 3F: Is there adequate participation of patients and/or their advocates in the peer review process?

While controversial to many, patients and their advocates have strongly articulated the view that they can provide a unique perspective to review of clinical research. Some argue that the presence of "non-scientists" erodes peer review, since a non-scientist is not a "peer." Others have argued that patients or their advocates have unique insight into certain aspects of clinical research, including the feasibility of the proposed work -- particularly with regard to recruitment issues and human subject protections.

The National Institute of Mental Health (NIMH) has piloted the use of public reviewers as full voting members on review committees since 2004. Public reviewers are members of the mental health community; for example, mental health consumers, family members, advocates, educators, and others. NIMH has found that these reviewers enhance the review process. For more information, see *Appendix I, Previous and Ongoing Peer Review Experiments*.

Recommended Action: Pilot the wider use of patients and/or their advocates on reviews of clinical research.

Challenge 3G: There is a need to increase review focus on potential impact, past investigator accomplishment, and innovation, to reduce emphasis on routine methodology.

There is general consensus that the review process has become overly dependent on routine methodological detail, at the expense of considering overall impact to science and health. A higher level of discussion may restore review to an honorific rather than onerous process. Recognizing impact and innovation as distinct review criteria may raise the excitement level of review.

To reiterate an important point, it is likely that several key actions would need to be implemented simultaneously to maximize a reviewer's effectiveness: i) application length must be significantly reduced; ii) the applications must be more focused on their potential impact, the accomplishment of the investigator(s) and the uniqueness or innovation of the work proposed; iii), the review itself must be shortened significantly; and iv), a reviewer must be empowered to explicitly rank all the applications reviewed in the panel in which he or she participates.

By placing emphasis on a specific "impact" section in all applications, the opportunity for multiple people to review just the impact statements of applications may be enhanced. In this manner, the Not Recommended for Resubmission (NRR) decision could be made for those applications lacking requisite impact, and initial ranking could be performed on the basis of ranking alone. Similarly, if a particularly Request for Applications required emphasis on the innovation or uniqueness of applications, a second ranking could be performed using this parameter.

Goal: Ensure the best use of charter reviewer members' time and expertise.

Recommended Action: Shorten summary statements by focusing solely on the merit of science as presented.

(see also Challenge 1, Reducing Administrative Burden on Applicants, Reviewers, and NIH Staff)

Recommended Action: Shorten the length of the application and align it to specific review elements.

(see also Challenge 1, Reducing Administrative Burden on Applicants, Reviewers, and NIH Staff)

Recommended Action: Have charter review members explicitly rank applications at the conclusion of a study section meeting.

(see also Challenge 2, Enhancing the Rating System)

Challenge 4

Optimizing Support at Different Career Stages

As illustrated previously in this document, the administrative burden on all investigators, who are writing ever more grant applications, and the NIH, which is reviewing ever more applications, is becoming untenable.

Supporting early-career investigators rose as a top challenge during the diagnostic phase of the peer review self-study, and it has been the top priority of the NIH Director for many years. However, there is also a need to enable greater productivity of highly accomplished NIH investigators with less administrative burden to applicants and reviewers.

***Challenge 4A*: Early-career investigators encounter lower success rates at every stage of new (type 1) R01 application submissions.**

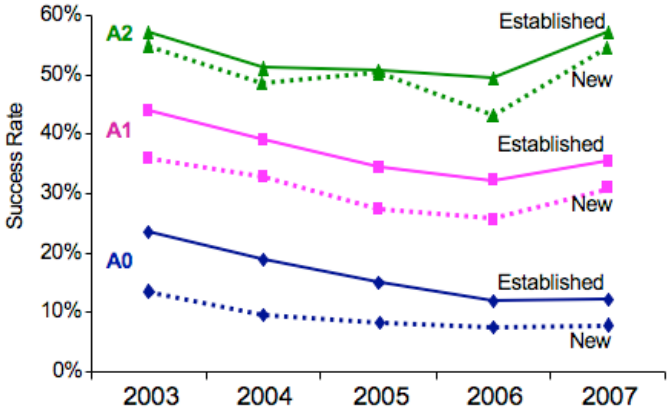


Figure 13: Comparison of success rates for early-career and established investigators. *Source: OER, Division of Information Services*

The success rate for early career investigators has decreased substantially since 1998 (Table 4) before taking an upturn in FY 2007, as a result of the proactive NIH commitment to fund a targeted number of early career investigators that year. The number of early career investigators funded in FY 2006 was a strikingly low 16.7 percent, as was the corresponding success rate for early career investigator applications, which in FY 2006 was 14.8 percent. In comparison, the funding rate for new (type 1) submissions for previously-funded NIH applicants was 21.3 percent in FY 2006, and the success rate that year for those applications was 17.5 percent. These data suggests that in difficult budget times, it is a significant challenge for a early career investigator to win their first R01. Furthermore, success rate numbers reflect multiple amended application submissions, and early career investigators may be especially disadvantaged (Figure 13). There has been a significant shift over the past decade from funding grants at their initial submission (A0) to funding grants at their second amended submission. Reviewers may be compelled to "mentor" early career investigators and demand improvements to their grant applications at a higher rate than they might demand of more established investigators.

Table 4. Funding rate for first-time R01-equivalent awardees. *Source: OER, Division of Information Services*

	First-Time Awardees		
FY	**Applicants**	**Awardees**	**Funding Rate**
1998	6,171	1,518	24.6 %
2000	6,752	1,612	23.9 %
2002	6,868	1,586	23.1 %
2004	8,155	1,539	18.9 %
2006	8,183	1,363	16.7 %
2007	7,758	1,596	20.6 %

The percentage of competing grants awarded to early-career investigators has been steadily dropping since 1989, and during the recent doubling of the NIH budget, the NIH did not create a vast new cohort of early-career investigator awardees. Rather, the bulk of the increase of the NIH budget went to established investigators.

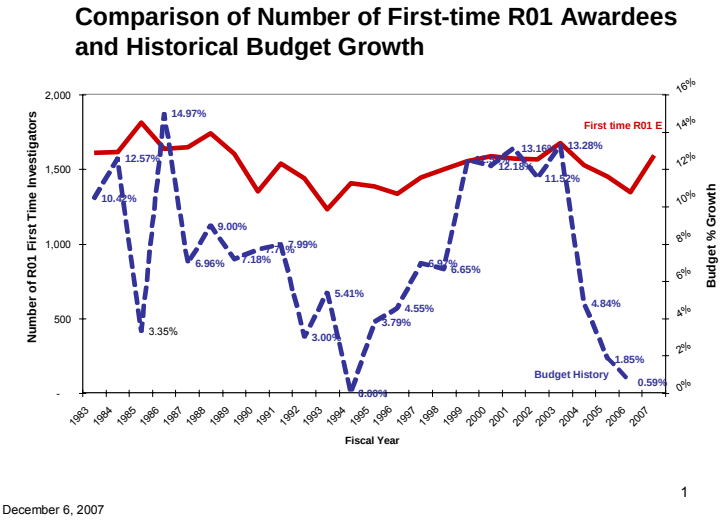


Figure 14: Comparison of the number of first-time awardees with historical budget growth. *Source: OER, Division of Information Services*

In fact, although growth of the NIH budget has fluctuated wildly over the past decades, the number of early career investigators supported by the NIH has remained roughly constant, hovering around 1500 since the early 1980s (Figure 14). A number of initiatives aimed at helping early career investigators have come and gone during this time period; it is useful to note that none have had a dramatic impact on increasing the number of early career investigators . This could be because the initiatives were ineffective, and/or that the NIH funding system somehow maintains its homeostasis in the face of a wide range of stimuli.

NIH must also consider the importance of growing the nation's research capacity, as well as preparing for the time when the "baby boom" generation retires. Careful modeling of the system, taking into account demographic and actuarial variables, should help guide the NIH in considering these issues.

(see also Challenge 6, Reducing the Stress on the Support System of Science)

<u>Challenge 4B</u>: The average age of early-career investigators has increased.

A related concern regarding early-career investigators is that the average age of the first R01-equivalent award[11] increased from 37 years in 1980 to 42 years today (Figure 15).

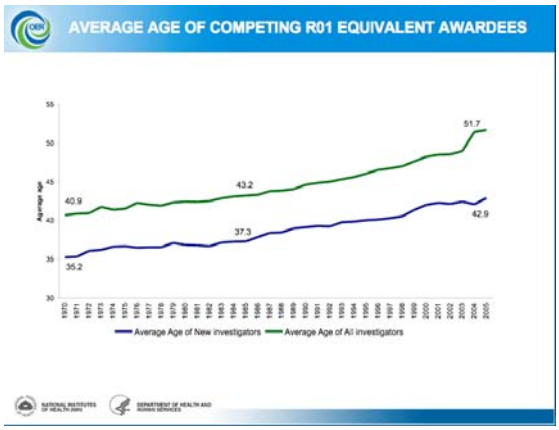


Figure 15. Comparison of age of first R01 for new and all investigators. *Source: OER, Division of Information Services*

<u>Challenge 4C</u>: There is an increasing gap between principal investigator appointment and first research project grant (RPG)[12].

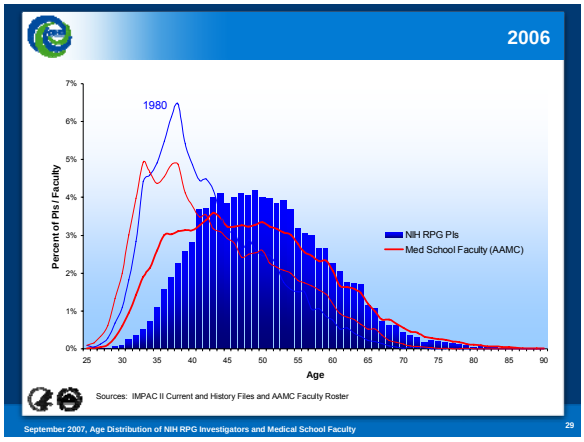


Figure 16. There is an increasing gap between PI appointment and first RPG . *Source: OER, Division of Information Services*

Part of this "graying" of the early-career investigator is due to a similar increase in age of first appointment to faculty position (from an average of 34 years in 1980 to an average of 39 years today). However, data (Figure 16) suggest that averages do not tell the whole story; there is a growing gap between age of first appointment and first research support. They are increasingly concerned about exhausting their startup package before they secure NIH funding. This tension was clearly evident in the peer review self-study RFI responses. These data also illustrate another startling point: the NIH is funding significantly more investigators over the age of 70 than under the age of 30.

[11] R01-equivalent grant = R01, R23, R29, R37
[12] RPG = R01, R23, R29, R37, DP1, P01, P42, PN1, R03, R15, R21, R22, R23, R33, R34, R35, R36, R37, R55, R56, UC1, UC7, U01, U19

***Goal*: Early-career investigators should, at a minimum, be on par with established principal investigators, in application success rates.**

There are a number of approaches that could be used to achieve this goal. One approach would be a unique award for early-career investigators. The potential advantage of doing so is that the early-career investigators would compete with each other for funding and not with established investigators. However, two programs, the R23 and the R29 (FIRST) award were discontinued by the NIH when it was determined that awardees of these mechanisms were less likely to secure subsequent R01 funding than early-career investigators starting with an R01 (Figure 17).

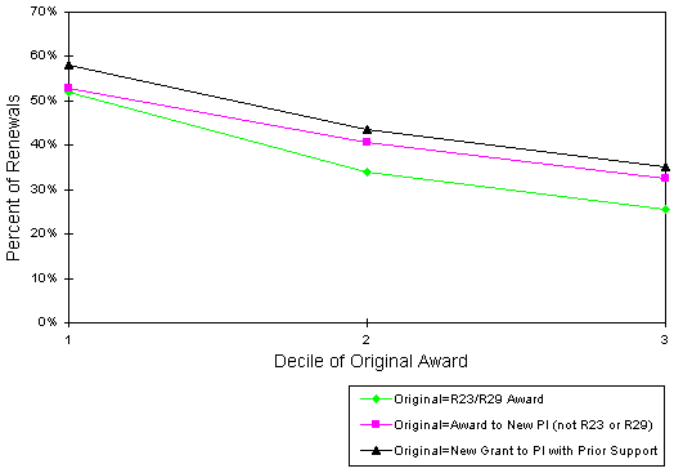


Figure 17. R23/R29 awardees are less likely to secure subsequent R01 funding. *Source: 1996 Report of the Working Group on New Investigators*

It was concluded that the R29 mechanism was under-resourced, and in the eyes of many, was perceived as less "impressive" than the R01 mechanism.

A variant of this approach would be to create a unique award for early-career investigators that is highly selective and would thus be awarded to only a small number of early-career investigators. The review of this award would emphasize creativity and innovation or uniqueness. The NIH Director's New Innovator Award program was initiated in 2007 to meet this need and has been received favorably (27).

A third approach would be to continue to administratively fund more R01s from early-career investigators. This approach has the advantage of providing early-career investigators independence as early in their careers as possible.

A fourth approach would be to review R01 applications from early-career investigators separately, either within unique study sections populated by generalists (which has been suggested by some to be a way to enhance risk-taking and greater innovation/uniqueness by applicants), or within the standard study sections. Note that the NIH Director's New Innovator Award uses generalists for reviewing proposals from early-career investigators.

Recommended Action: Continue to fund more R01s for early-career investigators.

This approach would not require any specific operational change, other than setting an internal "target" for NIH to fund a higher number of early career investigators. It expands the effort made in FY 2007 based on the historic 5-year NIH average (Table 5). This strategy increased the funding rate of early career investigators to 20.6 percent in FY 2007; it also closed the gap in success rates between early career and established investigators to 1.2 percent.

Table 5: Number of early-career investigators funded by NIH (FY 2002- FY 2007). Source: *NIH Office of Budget/OER, Division of Information Services*

	Number of first time awardees						FY 2007 Operating				
IC	**2002**	**2003**	**2004**	**2005**	**2006**	**FY 2002 - FY 2006**	**FY 2007 Operating Plan**	**FY 2007 Outcome**	**Made Goal? (√)**	**Number of Early-Career Investigator Applicants**	**2007 Early-Career Investigator Funding Rate**
NIH total TOTAL	**1,578**	**1,683**	**1,532**	**1,458**	**1,363**	**1,523**		**1,596**	√	**7,759**	**20.6%**

An advantage of this approach is that the early-career investigator receives an R01 grant. It creates no additional mechanism and would not add any administrative burden. A potential disadvantage to this approach is that in FY 2007, it created tension because the early-career investigator payline was significantly higher than the payline for more established investigators in many ICs. Study sections may adapt to this strategy and begin to adjust their scores in an attempt to "compensate" for what they perceive as inequity to more established investigators.

Furthermore, it is not known whether a 5-year average provides the best target for the number of early-career investigators for the NIH to support. OER and the NIH Office of Budget are modeling the NIH workforce with the goal of identifying a more evidence-based target in the future.

With the "reach" funding of early-career investigators in 2007, the NIH now has a rich data set to evaluate this approach. It would be interesting to compare those investigators that would not have been funded without the need to reach the target number of early-career investigators and the previous year's top scorers. Examining the productivity of both groups in 10 or so years would be an important exercise. This evaluation would begin to address whether study sections are evaluating early-career investigator applications appropriately, or if the applications are overly penalized for lack of preliminary data, less skill in grant writing, or lack of principal investigator "name recognition."

Recommended Action: Pilot the ranking of early-career investigators against each other.

This approach would allow early-career investigators to be reviewed in the study section most appropriate to their research focus, and yet by percentiling them separately, the applications from early-career investigators would be ranked against other early-career investigators, rather than against established investigators. The advantage of this approach is that it requires minimal administrative change, and would ensure that early-career investigator applications are reviewed in the most appropriate scientific context. However, it is unclear whether uneven review could be a problem since different ICs have used varying approaches to support early-career investigators.

Recommended Action: Pilot the separate review, by generalists, of early-career investigators, to enhance innovation and risk-taking by applicants.

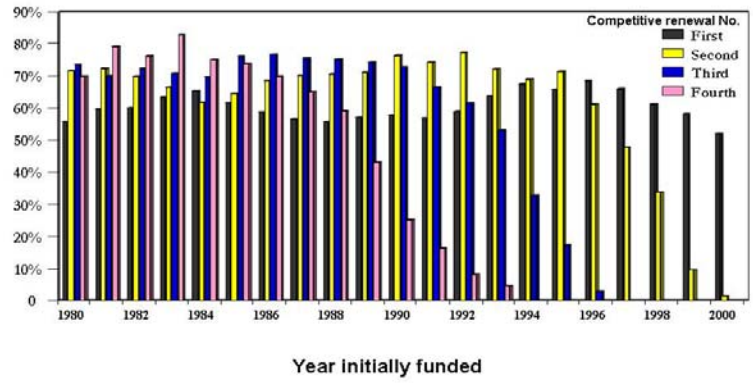
Given the enthusiasm many feel for supporting early-career investigators, it is possible that it would be easy to recruit excellent scientists to sit on these study sections, potentially allowing the NIH to recruit "generalists" or broad thinkers that could possibly assess the potential impact of new ideas free of worrying about technical minutiae. A possible disadvantage of having only generalists in these study sections is that it might be difficult to get appropriate scientific expertise to properly assess the scientific merit of the proposals.

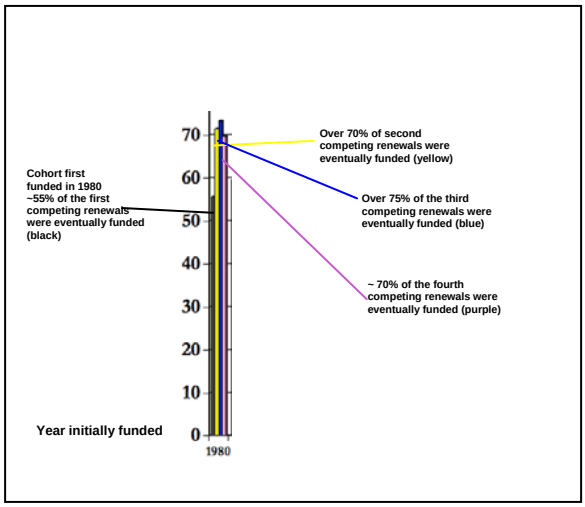
Recommended Action: The "environment" rating criterion for early-career investigators should take into account institutional commitment.

The NIH's investment in an early-career investigator is likely to be more lasting and enriched if his or her parent institution is committed to retaining the researcher as a faculty member. Institutions that commit resources to young faculty may, in turn, feel compelled to assure their chances at achieving scientific independence and career success.

Challenge 4D: There is a need to enable greater productivity of highly accomplished NIH investigators, with less administrative burden to applicants and reviewers.

Accomplished investigators have a strong track record. More than 70 percent of all competing renewals are eventually funded (Figures 18,19). The apparent "drop-off" observed in later years simply reflects the fact that these cohorts have not yet reached the time when they are eligible to re-compete for funding. Further, an analysis of NIH R37 awardees demonstrates that they too are highly successful in obtaining an R01 following the end of their MERIT/Javits award (Figure 20).





Figures 18 , 19. Percent of projects eventually funded (over time, top) as a function of competing renewal number. *Source: OER, Division of Information Services*

Many argue that the best predictor of future success is past performance. Some contest that the consideration of past performance places investigators earlier in their careers at a disadvantage, but even the newest investigators have a track record—indeed, it is on this basis that they are recruited to faculty positions. Nevertheless, more attention should be paid to the past accomplishments of the NIH's strongest investigators. By acknowledging the unique talents and contributions of these individuals, the NIH should be able to enable their continued progress with a lessened administrative burden for applicants, reviewers, and the NIH.

<u>Goal</u>: Enable greater productivity of highly accomplished NIH investigators, with less administrative burden to applicants and reviewers.

Recommended Action: Refine the NIH MERIT/Javits/NIH Director's Pioneer Award mechanisms to add a commitment to serve on a study section (if asked).

(see also Challenge 3, Enhancing Review and Reviewer Quality)

Though similar to current awards currently employed by many ICs, the modified award would carry with it greater opportunity and flexibility for the investigator. For illustrative purposes only, one option has the following characteristics: i) investigators would be invited to apply, rather than be administratively chosen for this award; ii) there would be trans-NIH consensus on the structure of the award; iii) a minimum of 51 percent effort would be required; and iv) awardees would be required to serve on a study section, if asked, as a condition of the grant award. Because it is expected that a relatively small number of investigators would qualify for this program, this program would likely comprise a small proportion of the total NIH investigator pool.

In this proposed scheme, a greater emphasis would be paid to past accomplishment as ascertained by a summary of the applicant's five most significant publications, the five most recent publications, and the five publications most germane to his or her proposed work. This award (for up to 10 years of funding) could offer investigators with a strong track record, and high likelihood of continued success, freedom from both the time of writing grant applications and the potentially repressive experience of undergoing review.

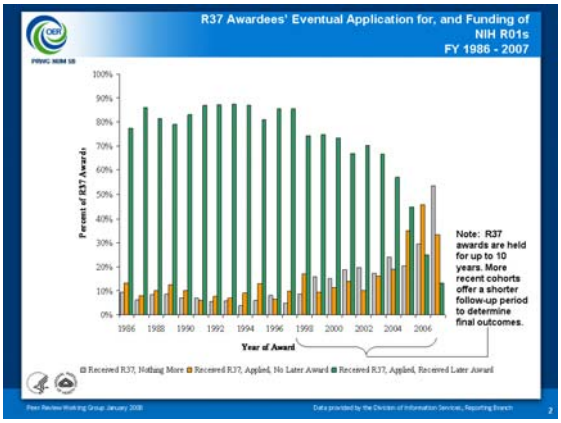


Figure 20. MERIT awardees have high success rates in securing subsequent R01 funding. *Source: OER, Division of Information Services*

Experience with the current MERIT award suggests that this mechanism would not be detrimental to the success of the research programs of the awardees. MERIT awardees are very successful when applying for subsequent R01 (and other RPG) funding upon completion of the term of the MERIT award (Figure 20). The advantages of stable funding and the option of being reviewed to a larger extent on past accomplishment should make this a very attractive mechanism for the NIH's most accomplished investigators.

Disadvantages of this approach include the fact that "cherry-picking" scientists could have the unintended effect of favoritism and potentially, elitism. The community may react negatively to the idea that, in times of constrained budgets, the most accomplished scientists are being "protected" by longer-term awards. However, MERIT awards comprise a relatively small proportion of all NIH R01 grants (Figure 21): approximately 3 percent by number and approximately 4 percent by dollar amount.

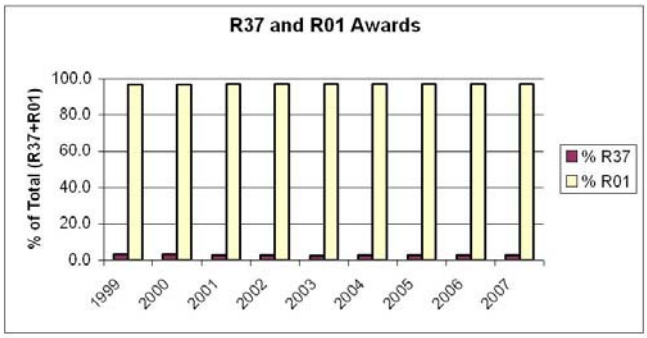


Figure 21. Proportion of NIH R37 (MERIT, Javits) awards compared to NIH R01 awards. *Source: OER, Division of Information Services*

<u>Challenge 4E</u>: The NIH needs to better understand the career needs of research associates/staff scientists.

(see also Challenge 6, Reducing the Stress on the Support System of Science)

The scale and complexity of today's biomedical research problems demand that scientists move beyond the confines of their individual disciplines and explore new organizational models for team science. The NIH wants to stimulate new ways of combining skills and disciplines in the physical, biological, and social sciences to realize the great promise of 21st-century medical research. Shared core facilities, established a research institutions, have long been viewed as essential tools to enable a number of established and independently funded investigators and research teams to have the opportunity to enhance their collective productivity to a greater degree than would be possible from the separate projects.

As was noted in the 2005 "Bridges to Independence" National Research Council report (28), very few postdoctoral scholars obtain a tenure-track position in academia, and the number of tenure-track positions has been constant over time. However, many postdocs still remain in the academic sector; the number of postdocs who pursue other academic positions has dramatically increased, from approximately 1,000 in 1985 to nearly 17,000 in 2001 (29). Some of these individuals work as part of a team in large-scale science projects, and it is appropriate to fund them through grants awarded for these projects. At research universities, faculty-level jobs lacking the possibility of tenure have risen from 55 percent of new hires in 1989 to 70 percent today (30).

Recommended Action: Develop a census of research associates/staff scientists as an initial step towards exploring approaches to providing more stable support for these individuals.

(see also Challenge 6, Reducing the Stress on the Support System of Science)

Challenge 5

Optimizing Support for Different Types of Science

Diverse types and approaches of science—clinical, basic, transformative, incremental, data collection, tool-building, interdisciplinary—are needed for progress toward improving human health. The NIH must strike a balance among these areas to fund the most meritorious science and address public health need.

Challenge 5A: There is a need to seek out and support the most transformative research ideas.

During the consultation phase of the peer review self-study, significant concern was raised that the current peer review system discourages creativity and innovation, while favoring incremental discoveries and tolerating repetition. Many noted that a disproportionate amount of an application's content is centered on preliminary data and methods, which can be a roadblock for the evaluation of innovative science. As indicated earlier, the term innovative is notably difficult to define with precision. Throughout this report, the term is intended to refer to research that cuts new ground from a conceptual or technical perspective, or is strikingly distinct or even unique compared to other ongoing research. The NIH clearly sees the need to strike a balance between fostering paradigm-shifting research and incremental research, because both are clearly needed to move science forward.

Goal: To provide clear opportunities for applications proposing transformative research

Recommended Action: Use the NIH Director's Pioneer, NIH Director's New Innovator, and the Exceptional, Unconventional Research Enabling Knowledge Acceleration (EUREKA) award programs as starting points to develop a path to invite, identify, and support transformative research, expanding the number of awards to a minimum of 1 percent of all R01-like awards.

Funding proposals with high potential impact submitted by accomplished investigators is often considered a way to support high-risk, high reward research. The NIH Director's Pioneer Award Program (begun in 2004) and the NIH Director's New Innovator Award Program (begun in 2007) followed this principle and were established with the goal of promoting innovative research among established and early-career investigators, respectively. Both programs are meant to support exceptionally creative investigators proposing highly innovative, and possibly high-risk, projects. The novel peer review processes for these programs were designed to address the perception that applicants do not propose their most innovative and unique ideas in R01 applications and that reviewers on traditional study sections tend to be conservative and do not give the best scores to highly innovative or risky research projects.

The EUREKA award program[13] seeks exceptionally innovative research on novel hypotheses or difficult problems, solutions to which would have an extremely high impact in biomedical or bio-behavioral research. This program features a unique application structure and review process.

Using these programs, the NIH could create a discrete and highly visible transformative path that would comprise at least 1 percent of all NIH R01-like awards. Such awards would provide substantial support for truly transformative ideas. They would employ a distinct application and review process that features extremely rigorous standards. Inherent to this approach is the realization that extremely transformative ideas have inherently high failure rates and may be slow to develop or meet conventional metrics of success, such as publications or widespread acceptance. Thus, an important feature of awards for transformative research is the trust shown by the NIH for its awardees in this program.

***Challenge 5B*: There are differences in success rates for applications proposing clinical research than applications not proposing clinical research.**

Many stakeholders raised the issue of clinical science success in the NIH review process. There is a perception within the clinical research community that basic research is favored in the peer review process. Part of this is fueled by data that reveals that the locus of review appears to affect the success rate of clinical research applications (Figure 22).

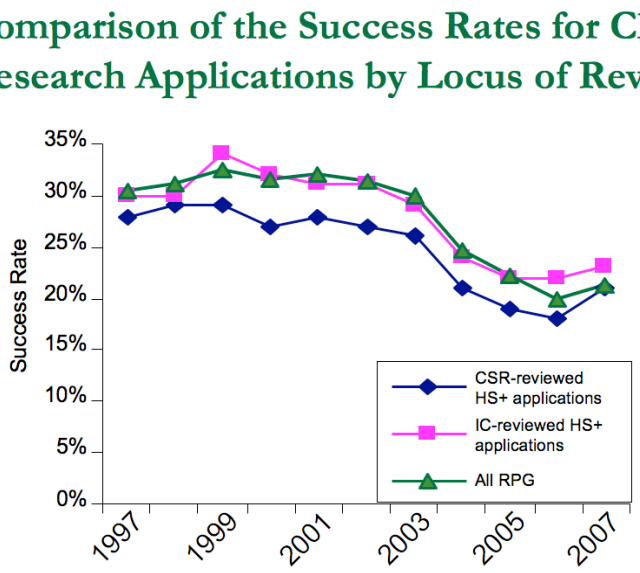


Figure 22. The locus of review affects clinical research application success rate. *Source: CSR*

***Challenge 5C:* For applications reviewed in CSR, "non-clinical" R01s fare better than "clinical[14]" R01s, in part, because clinical research applicants appear less likely to send in amended type 1 (new) submissions or type 2 (competing continuation) submissions.**

[13]EUREKA program FAQ: http://www.nigms.nih.gov/Research/Application/EurekaFAQs.htm
[14] Clinical: "Human subjects-positive"

Assuming no overall difference in the quality between clinical and basic research, if the two application types were reviewed in the same manner, then the respective lines in Figure 23 would be expected to lie on diagonal such that, for example, 20 percent of the clinical and basic research applications would score in the 20th percentile. Instead, the data show that non-clinical research tends to score above the reference line, while clinical research scores below it. For example, 22.1 percent of the non-clinical proposals and 17.6 percent of clinical proposals each score in the 20th percentile.

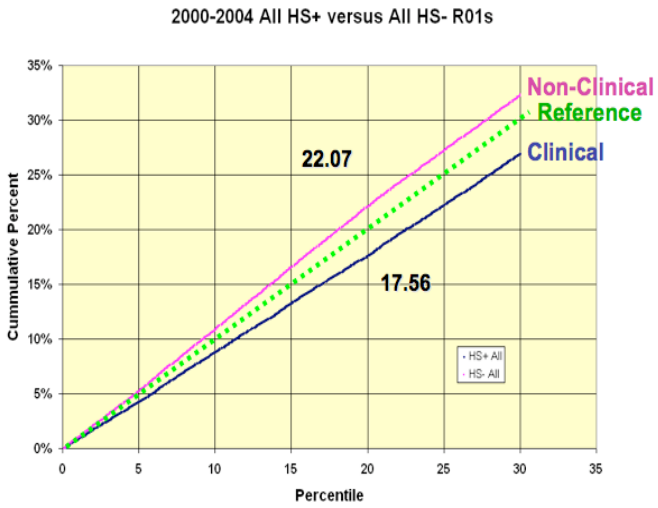


Figure 23. For applications reviewed in CSR, "non-clinical" R01s fare better than "clinical" R01s. *Source: CSR*

The differences observed in CSR appear to be related to differences in the submission patterns between clinical and non-clinical investigators. Table 6 demonstrates that clinical investigators are less likely to submit amended type 1 applications or type 2 applications--both categories that fare better in terms of success rate.

Proportional Distribution of Different Types of R01 Applications			
	HS+	**HS-**	**Difference (%)**
Type1NewA0	**24.59%**	**20.20%**	**4.39%**
Type1NewA1	9.20%	7.68%	1.52%
Type1NewA2	2.30%	1.94%	0.36%
			0.00%
Type1ExpA0	**28.12%**	**27.58%**	**0.54%**
Type1ExpA1	12.57%	11.55%	1.02%
Type1ExpA2	3.62%	3.40%	0.22%
			0.00%
Type2A0	**11.71%**	**17.11%**	**-5.40%**
Type2A1	5.95%	7.90%	-1.95%
Type2A2	1.96%	2.65%	-0.69%
Total	100.00%	100.00%	

Table 6. Clinical research (HS+) applications appear less likely to submit amended Type 1or Type 2 applications. ***Source: CSR***

A second consideration surfaced from CSR's analysis of clinical science review. Study-section members are required to factor human subject concerns into their assignment of a priority score, since this is an important aspect of the study design, which needs to be considered in review. It has emerged that human subject protection concerns were raised in 12.9 percent of all clinical applications, and these applications fared particularly poorly in review. Within the subset of applications where reviewers identified human subject concerns; only 9.96 percent received a percentile rank of 20.0 or better (Figure 24). Human subject concerns with an initial application are not confined to early-career investigators, although this population does represent the largest group. Even some experienced, funded clinical researchers proposing continuation of their research have difficulty in documenting adequate protections for human subjects. In practice, most applicants address these human subject concerns in their revised application. No application may be funded until all human subjects concerns are resolved.

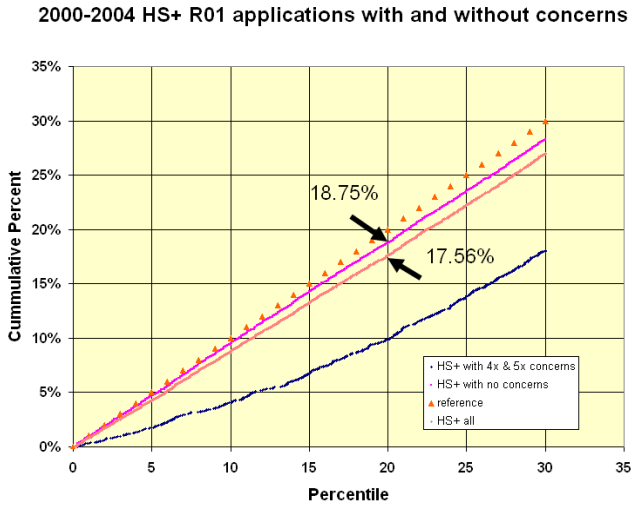


Figure 24. Concerns about human subjects protections documentation influence review outcome for clinical (HS+) research applications. *Source: CSR*

Goal*: *To ensure optimal review of clinical research

Recommended Action: Determine the underlying causes of submission patterns and results in CSR and IC panels and consider corrective actions if needed.

The reasons for the different behavior of clinical and non-clinical investigators in submitting competitive renewals are not clearly understood and they should be studied further. It is possible that many more of the clinical applications have clear endpoints. Once these endpoints are reached and the study is over, the investigator may move on to another defined project. Basic research, by its nature, even when very successful, often moves the research project on to the next question, building upon the knowledge obtained, and leads the investigator more naturally, toward a competing continuation application. Another possibility is that clinical R01s reviewed within CSR have particular characteristics that are different from the clinical R01s reviewed by ICs obtained, for example, in response to Requests for Applications and Program Announcements. Yet

another possibility to consider is that submission rates for competing continuation applications may be lower because clinical investigators stop doing research altogether, or if they obtain funding from other institutions or industry.

With respect to the issue of addressing human subject protection concerns within the review process, data suggest that training for early-career clinical investigators (as well as for established ones, but to a lesser extent) in appropriate research design and protections for human subjects would likely help the review outcomes of clinical proposals.

Recommended Action: Ensure participation of adequate numbers of clinician scientists by providing more flexible options for service.

(see also Challenge 3, Enhancing Review and Reviewer Quality)

Recommended Action: Continue to pilot the use of patients and/or their advocates in the review of clinical research.

(see also Challenge 3, Enhancing Review and Reviewer Quality)

Challenge 5D: Interdisciplinary research needs a space to be reviewed and supported.

Stakeholders have reported that interdisciplinary research is often reviewed as "unfocused and overly ambitious in nature." The NIH Roadmap for Medical Research has piloted a number of initiatives to enhance the review and support of interdisciplinary research (http://nihroadmap.nih.gov/interdisciplinary), including new mechanisms (*e.g.*, interdisciplinary research consortia, T90/R90 training programs). In addition, the NIH has now formally recognized multiple principal investigators as a means to facilitate multi- and interdisciplinary research teams.

Goal: To ensure the optimal review and support of interdisciplinary research

Recommended Action: The NIH should analyze applications that are interdisciplinary in nature with respect to: i) referral patterns for review; ii) assignment for secondary review and funding consideration; and iii) success rates.

Recommended Action: Employ an editorial board model for the review of interdisciplinary research that includes content experts, "big picture" thinkers, and "interpreters."

Recommended Action: Enhance trans-NIH approaches to provide support space for highly meritorious interdisciplinary research.

Challenge 6

Reducing Stress on the Support System of Science

Regardless of the numerous and complex issues that form the contextual landscape surrounding the stresses on the system used to support research in this country, resources will always be finite in nature. The NIH must continue to guide the distribution of these resources through careful and transparent prioritization in concert with the NIH's stakeholders.

Challenge 6A: The NIH funding system has finite resources.

Due in part to a flattened budget and in part to an increase in applications, NIH funding success rates have declined (Figure 25).

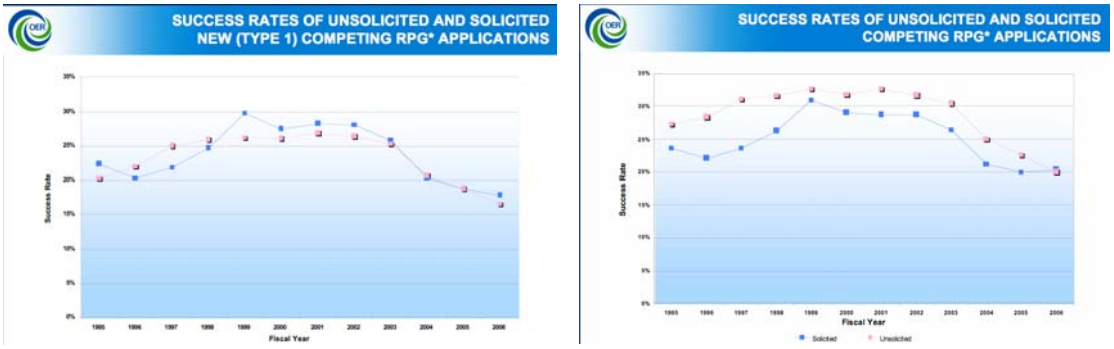


Figure 25. NIH funding success rates for new (left) and competing continuation (right) applications. *Source: OER, Division of Information Services*

Approximately 60 percent of NIH-supported investigators have under $400,000 in total direct costs (Figure 26).

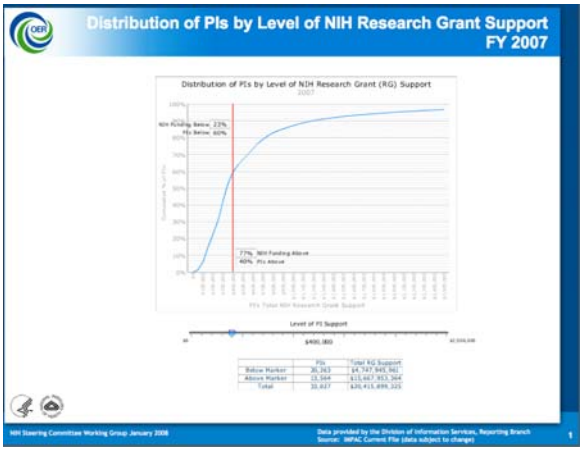


Figure 26. Distribution of NIH-funded investigators by level of grant support. *Source: OER, Division of Information Services*

There is an increasing pressure for all investigators, whether tenured, on tenure-track or not, to recover salary support from grants. With modular grants capped (with

approximately 60 percent of competing modular applications now at the "cap") and the policy to not increase overall average cost of grants, investigators may seek multiple awards to obtain the resources to support their research efforts. In addition, investigators may prefer multiple awards that have "staggered" end-dates, to ensure some continuity of support even when one grant might be lost. Many have argued that mechanisms such as the R03 and R21 are being used as "mini" R01s--in some instances, these may used as "stepping stones" for investigators gathering needed preliminary data for an R01 application. In other cases, these smaller awards may be used to make up resource shortfalls. At least one Institute at NIH (NIGMS) has stopped using the R21 mechanism, and several ICs limit R03s to early-career investigators.

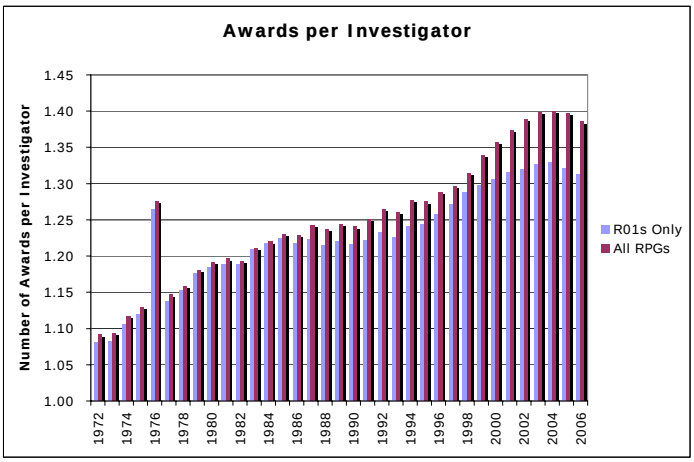


Figure 27. Number of NIH awards per investigator (R01-blue, all RPGs-red). Note: 1976 had a 15-month fiscal year, allowing an extra Council cycle to be captured in the data for that year. *Source: OER, Division of Information Services*

Although the number of research project grants per investigator has been steadily rising (Figure 27), less than 2.5 percent of NIH-supported investigators have four or more research grants[15] (Figure 28). These data suggest that a relatively small proportion of NIH-funded investigators have multiple R01 grants, and that more than half of the NIH-funded investigators have two or less R01 grants.

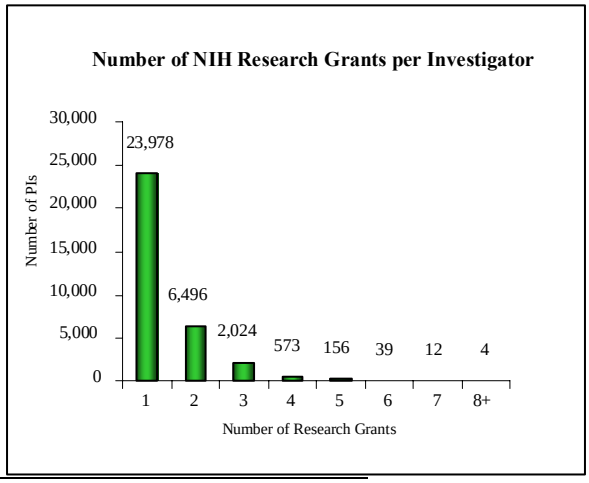
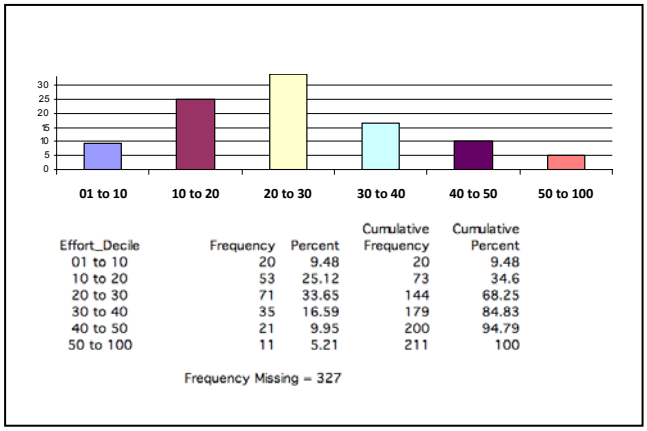


Figure 28. Number of NIH research grants per investigator. *Source: OER, Division of Information Services*

[15] Research Grant – awards made to Research Centers (G12, M01, P20, P30, P40, P41, P50, P51, P60, PL1, R07, U30, U40, U41, U42, U50 and U54), Research Projects (DP1, DP2, P01, P42, PN1, R00, R01, R03, R15, R21, R22, R23, R29, R33, R34, R35, R36, R37, R55, R56, RL1, RL2, RL5, RL9, U01, U19, U34, UC1, UC7), SBIR/STTR (R41, R42, R43, R44, U43, U44), Research Career (All K awards) and Other Research Grants (PN2, R13, R18, R24, R25, R90, RC1, S06, S10, S11, S21, S22, SC1, SC2, SC3, U10, U13, U18, U24, U2R, U45, U56, UH1, UL1).

Anecdotally, it has been reported that principal investigators apply for an application with a percent effort that is often greater than the percent actually implemented once an award is made. In part, this is driven by the uncertainty about which applications will be funded and when. In a pilot survey of 370 non-modular grants, 70 percent of principal investigators dedicated less than 30 percent effort to any given grant (Figure 29).



Effort_Decile	Frequency	Percent	Cumulative Frequency	Cumulative Percent
01 to 10	20	9.48	20	9.48
10 to 20	53	25.12	73	34.6
20 to 30	71	33.65	144	68.25
30 to 40	35	16.59	179	84.83
40 to 50	21	9.95	200	94.79
50 to 100	11	5.21	211	100

Frequency Missing = 327

Figure 29. Investigator percent effort dedicated to NIH grants. *Source: OER, Division of Information Services*

Two-thirds of NIH principal investigators have 50 percent or less in aggregate support (Figure 30).

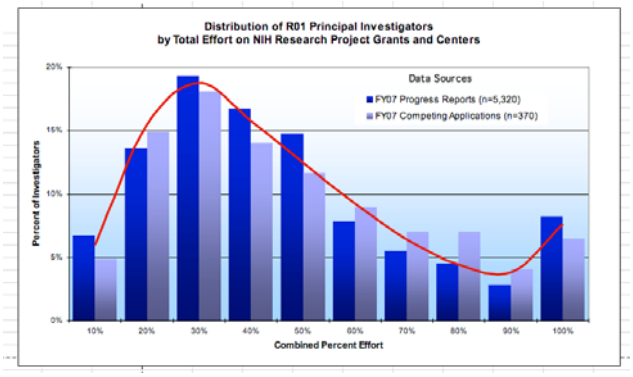


Figure 30. Distribution of NIH-funded investigators by percent effort on research project grants. *Source: OER, Division or Information Services*

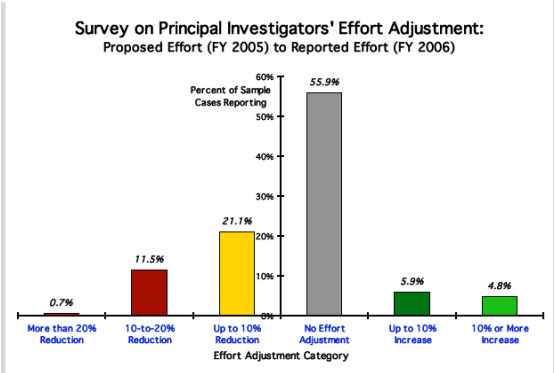


Figure 31. Adjustment of principal investigator percent effort after award. *Source: OER, Division of Information Services*

More than half (56 percent) of investigators do not adjust their effort as described on their original proposal. However, in one survey[16], 33 percent of investigators reduced their commitments, while only 11 percent of investigators increase their originally stated commitments (Figure 31).

Goal: To ensure the optimal use of NIH resources

Recommended Action: Require, in general, a minimum percent effort on research project grants.

It is recommended that principal investigators devote at least 20 percent effort per grant, unless they can provide an explicit justification to the NIH for a lower percent effort. All remaining participants should devote a minimum of 5 percent effort per grant, unless they can provide explicit justification to the NIH for a lower percent effort.

De facto, setting a minimum percentage effort would decrease the number of research grants any one investigator could hold than is currently possible. Such an action would also require an augmentation of current post-award accounting to ensure that final percent effort is appropriately captured for each investigator. To ensure that science does not suffer due to lack of adequate funds, the NIH would need to continue to allow applications to be "right-sized," *i.e.*, budget requests must be appropriately justified by scientific need. Note that the implementation of this recommendation would require clarity regarding the definition of percent effort to be used.

Challenge 6B: Universities continue to build additional research facilities, populated increasingly by people on "soft money," non-tenure track positions.

Universities, and in particular academic health centers, continue to build research buildings, populated increasingly by individuals dependent exclusively on grants for their salaries (soft money positions), to attract more grant dollars (31,32). Many have expressed concern that the current incentives in the NIH funding system are causing the research community to expand in ways that are not optimal and that cannot be sustained. Further, many have called for universities to take responsibility for a larger fraction of all faculty salaries. While institutions are built on different business models, there would be great value in beginning a serious dialogue between the NIH and all its stakeholders to explore the advantages and disadvantages of the current NIH model for salary support. As an initial step, the NIH must collect data on its investigators' aggregate percent effort and salary recovery.

There are now more Ph.D. scientists appointed in clinical departments of academic health centers than there are in basic science departments. The majority of these researchers are on soft money appointments. An increasing number of soft money positions is part of a broader trend observed for institutions of higher education in the United States. Between 1987 and 2003, the percentage of faculty who were tenured/tenure-track decreased by 15

[16] Data represents pilot survey of 370 non-modular grants

percent, a decrease of one out of every seven traditional tenure-eligible positions. Full- and part-time non-tenure track appointments accounted for three out of five faculty positions and approximately 75 percent of new hires (33). Many have questioned if there is sufficient regulatory control of this system: Clearly, the interdependence between universities, research institutes and other organizations engaged in the conduct of biomedical and behavioral research, and the NIH, must be subject to a continued analysis.

For example, what would be the effect on the overall future expansion rate if plans were made to limit the percentage of a principal investigator's salary that could be recovered? Should incentives (via indirect cost-rate rules, for example) be put into place to encourage a greater percentage of a principal investigator's salary to be supported by his or her institution?

Because institutions are built with different business models, any action of this nature would need to be phased in over time and in consultation with the NIH's many stakeholders.

***Goal*: Optimize the system used by the NIH to support principal investigators and other research personnel.**

Recommended Action: The NIH should analyze the incentives inherent within the NIH system of funding that have been driving the rapid expansion of the U.S. biomedical research system in recent years and explore with its stakeholders whether these incentives should be reduced or eliminated.

***Challenge 6C*: The number of tenure-track positions in academia, and scientist positions in all sectors, is straining to keep up with the number of postdoctoral fellows being trained.**

Demand for NIH support is driven in part by the size of the biomedical work force. In turn, the current workforce is determined in part by a need to sustain research efforts by ensuring an ample supply of talented and highly skilled workers in the form of graduate students and postdoctoral fellows. This has contributed to an ever-increasing lengthening of the scientific training period, resulting in significant delays in scientists achieving independence. Most of these trainees are supported on R01s.

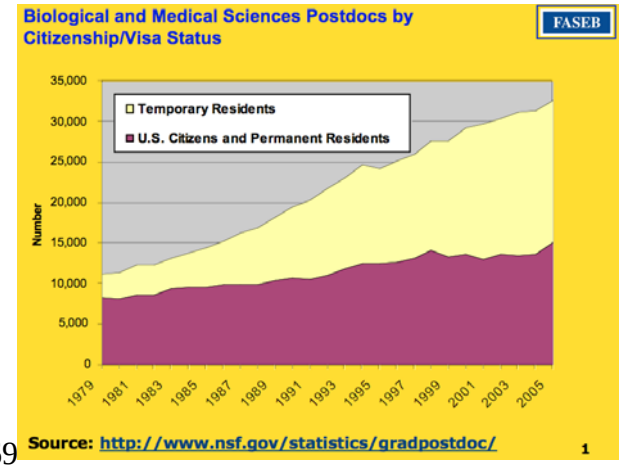


Figure 32. The ever-increasing number of postdocs without a concomitant increase in available faculty slots. *Source: FASEB*

While the demand for postdoctoral fellows continues to grow (Figure 32), the number of tenure-track positions in academia has not kept pace. For example, the total number of new (Ph.D.) faculty in U.S. medical schools continues to decline (Figure 33). Industry has been the fastest growing employment sector, but recently it appears that the expansion in this sector is also slowing.

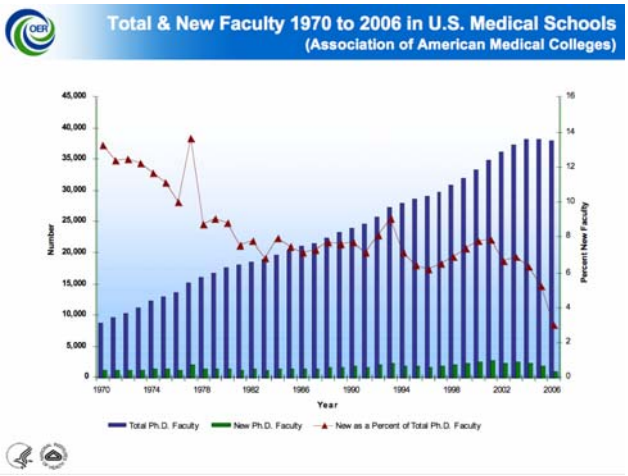


Figure 33. Total and new faculty 1970 to 2006 in U.S. medical schools. *Source: AAMC*

Recommended Action: Analyze the NIH contribution to the optimal biomedical workforce needs.

i) Evaluate the total number of graduate students and postdoctoral fellows being supported.

ii) Develop a census of research associates/staff scientists as an initial step towards exploring approaches to providing more stable support for these individuals.

Challenge 7

Meeting the Need for Continuous Review of NIH Peer Review

A cornerstone of the system employed by the NIH to support biomedical and behavioral research is the two-tiered peer review process. The current peer review self-study took on the goal of optimizing the efficiency and effectiveness of peer review, to ensure that the NIH will be able to continue to meet the needs of the research community and the public-at-large.

Challenge 7A: The biomedical and behavioral research enterprise is highly dynamic and peer review must evolve to keep pace.

Critical to the health of the NIH peer review system is ongoing evaluation. Thus, it is critical that the NIH capture appropriate baseline data from the current peer review self-study and develop appropriate measures and indicators for future monitoring efforts.

Goal: To assure the core values of peer review

Recommended Action: Mandate a periodic, data-driven, NIH-wide assessment of the peer review process.

Recommended Action: Capture appropriate current baseline data and develop new metrics to track key elements of the peer review system.

A series of predictors or benchmarks should be identified that will inform the outcomes of the recommendations in this report. These predictors or benchmarks should be made known to the community, and the data with appropriate analyses should be shared widely at appropriate intervals. Thus, the outcomes of any peer review pilots will be apparent to NIH's stakeholder communities, enabling them to engage in informed discussions about potential modifications. It should be noted that a decision not to change a current process also represents a decision that must be subject to testing; the NIH must validate the status quo with equal rigor as for any proposed changes.

Specific opportunities for data capture emerged as a result of the 2007-2008 peer review self-study. For example:

- Monitor reapplication and success rates for early-career investigators as their first independent grant come up for renewal.
- Continue to monitor the number of grant applications using different mechanisms over time.
- Monitor the use of the NRR checkbox and the behavior of the applications with regard to the submission of other proposals.

- Collect information from applicants and reviewers regarding their views of application length.

- Monitor the characteristics of chartered reviewers serving of study sections.

- Monitor the percent effort levels on individual research project grants as well as levels aggregated over from all NIH mechanisms.

- For clinical science, determine underlying causes of submission patterns and different results observed between CSR and IC panels. In addition, determine if an optimal number of clinician scientists serve on review panels.

- For interdisciplinary science, analyze applications that are IR in nature with respect to:
    - Referral patterns for review
    - Assignment for secondary review and funding consideration
    - Success rates

SUMMARY OF RECOMMENDED ACTIONS

The 2007-2008 NIH peer review self-study has underscored the need for the NIH to assure that the processes used to support science are fair, efficient, and effective. Key challenges, as depicted in Figure 34, were identified.

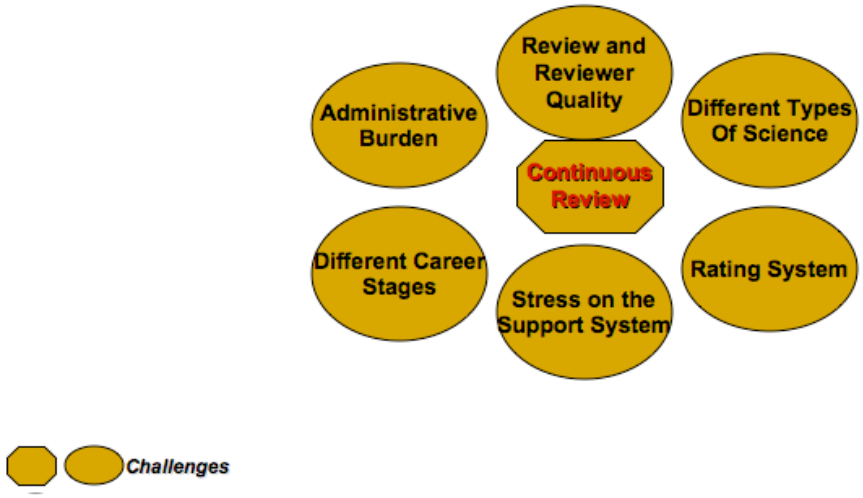


Figure 34. Challenges to enhancing the NIH peer review process.

Addressing the several key challenges to sustaining quality peer review in the modern biomedical and behavioral research climate have pointed to a set of overlapping solutions and benefits (Figures 35, 36). Vital to the success of optimizing peer review is that the NIH maintain the core values of peer review: scientific competence, fairness, timeliness, and integrity.

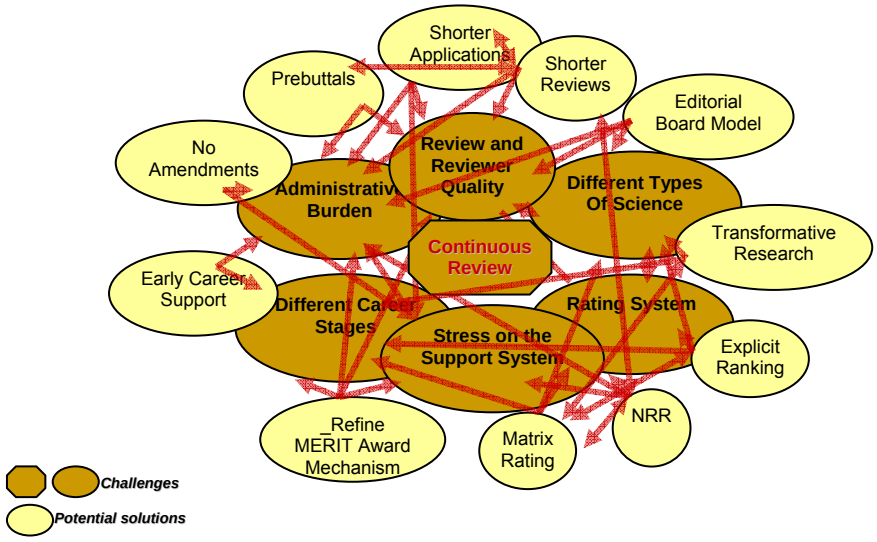


Figure 35. Linked challenges and solutions to enhancing the NIH peer review process.

The 2007-2008 peer review self-study determined that reducing administrative burden for applicants, reviewers, and NIH staff is a necessary component of enhancing peer review. Meeting the principal challenge of rising administrative burden requires reducing the number of applications that need to be submitted, shortening reviews by focusing solely on scientific merit as presented, and reducing the length of applications.

Another key focus is to assure that the rating system used to assess NIH applications is both accurate and sufficiently informational for decision making. Steps that can be taken to achieve this goal include rating multiple, explicit criteria individually, but providing an independent, overall score. Giving clear and unambiguous feedback to all applicants will be enhanced through the use of a "Not Recommended for Resubmission" (NRR) category, and shorter, criterion-aligned applications will streamline review while encouraging focus on innovation, uniqueness, and impact.

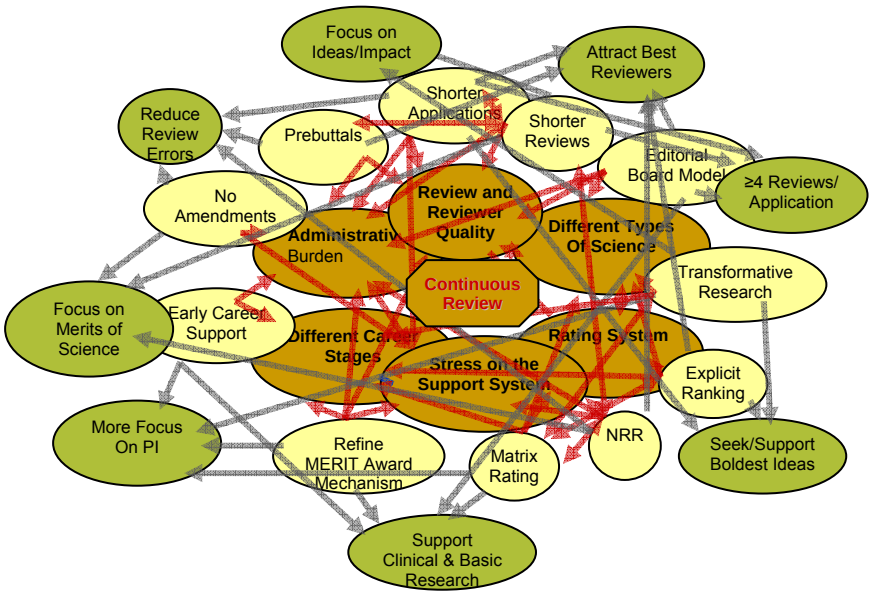


Figure 36. Combinatorial network showing outcomes achieved (green) by combining peer review enhancement solutions.

A linchpin of review quality is recruiting and retaining excellent reviewers. Addressing the larger problem of changing the culture of review can be achieved through reviewer incentives, refocusing review behavior on scientific merit as presented, and administrative actions that reduce burden (such as shorter applications and summary statements). Engaging more reviewers per application and throughout the review process will help to ensure review quality and consistency, as would enhanced reviewer training.

Another challenge facing the peer review system is the heterogeneity of review needs for different populations of applicants. The 2007-2008 peer review self study revealed that

reviewers struggle with dealing with both ends of the career continuum: investigators just starting out and those that are well-established. Continuing to fund more R01s for early-career investigators, and potentially applying specialized reviews for them, demonstrates the NIH's keen interest in protecting the pipeline. Highly accomplished investigators may be better served through retrospective evaluation and longer funding periods, and this could also benefit the NIH by coupling these actions to potential review service.

As has always been the case, diverse types and approaches of science are necessary to fulfill the NIH's mission to improve the nation's health, and peer review must accommodate the NIH's need to strike an appropriate balance among these. Using the current NIH Director's Pioneer, NIH Director's New Innovator, and Exceptional, Unconventional Research Enabling Knowledge Acceleration (EUREKA) Award programs, the NIH could increase its commitment to funding truly transformative ideas. It is also key that the NIH determines the underlying causes of clinical research application submission patterns that have been observed over time, and if necessary, implement corrective measures. Assuring that distinct modes of science, such as interdisciplinary research, receive fair and accurate review may entail tailoring review as needed.

To continue to guide the distribution of the NIH's finite resources, it is important to initiate a dialogue with stakeholder communities regarding the issue of salary support for investigators, recognizing the diversity of business models employed by applicant organizations. In addition, by requiring a minimum percent effort for investigators on research project grants, the NIH will ensure the optimal use of its resources.

Finally, it is critical that the NIH establish transparent and able mechanisms to continually assess the health of the peer review system. Ongoing, data-driven efforts must evaluate review outcomes and test the success of pilot programs. Key to attaining this goal is capturing appropriate current baseline data and developing new metrics to track key elements of the peer review system.

ACKNOWLEDGEMENTS

We acknowledge with our deepest thanks the truly outstanding efforts of our team: Amy Adams, Kerry Brink, Alison Davis, Vesna Kutlesic, and Jennifer Weisman. We are most grateful for the contributions made by Penny Burgoon and Stefano Bertuzzi who stepped in at key moments. Great appreciation is extended to members of the OER team that gathered and analyzed much of the data that appear in this draft report: Sally Rockey, Israel Lederhendler, and Jim Onken. We also thank members of CSR who also provided significant amounts of data. We are most grateful to the members of the ACD and SC WGs for their considerable efforts. Input from a number of IC Directors not directly involved with the project is gratefully acknowledged.

Dr. Jeremy Berg, Co-Chair, Steering Committee *Ad Hoc* Working Group on NIH Peer Review

Dr. Keith Yamamoto, Co-Chair, Working Group of the Advisory Committee to the NIH Director (ACD) on NIH Peer Review

Dr. Lawrence Tabak, Co-Chair, Steering Committee *Ad Hoc* Working Group on NIH Peer Review, Co-Chair, Working Group of the Advisory Committee to the NIH Director on NIH Peer Review

REFERENCES

1. Spier R. The history of the peer-review process. *Trends Biotechnol.* 2002; 20:357-8.

2. Al Kawi MZ. History of medical records and peer review. *Ann. Saudi. Med.* 1997;17:277–8.

3. Ajlouni KM, Al-Khalidi U. Medical records, patient outcome, and peer review in eleventh-century Arab medicine. *Ann. Saudi Med.* 1997;17:326–7.

4. Allen, EM. Early Years of NIH Research Grants. *NIH Alumni Association Newsletter* 1980; 2:6-8:

5. Recommendations For Change At The NIH'S Center For Scientific Review: Phase 1 Report: Panel on Scientific Boundaries for Review 2000: http://cms.csr.nih.gov/NewsandReports/ReorganizationActivitiesChannel/BackgroundUpdatesandTimeline/FinalPhase1Report.htm (Accessed February 27, 2008)

6. Scarpa T. Peer Review at NIH. *Science* 2006; 311:41.

7. Companario, JM. Peer review for journals as it stands today-Part 1. *Science Communication* 1998;19:181-211.

8. Companario, JM. Peer review for journals as it stands today-Part 2. *Science Communication* 1998;19:277-306.

9. Stehbens, WE. Basic philosophy and concepts underlying scientific peer review. *Medical Hypotheses* 1999;52:31-6.

10. Kotchen, TA,Lindquist, T, Malik K, Ehrenfeld, E. *JAMA* 2004 ;291 :836.

11. Pagano, M. American Idol and NIH grant review. *Cell* 2006;126, 637-8.

12. Munger, K. American Idol and the NIH grant review-Redux. *Cell* 2006;127:661-2.

13. Miller G, Couzin J. NIH BUDGET: Peer Review Under Stress. *Science* 2007;316:358-9.

14. Kohane IS, Altman RB. Proc AMIA Symposium, 2000, 433-7: http://bmir.stanford.edu/file_asset/index.php/933/SMI-2001-0994.pdf (Accessed February 28, 2008)

15. Dowdy, SF. The anonymous American Idol manuscript reviewer. *Cell* 2006;127:662-3.

16. Dinov, ID. Grant review: American Idol or Big Brother? *Cell* 2006;127:663-4.

17. Bielski A, Harris R, Gillis N. Summary Report of Comments Received on NIH System to Support Biomedical and Behavioral Research and Peer Review: http://enhancing-peer-review.nih.gov/meetings/Peer_Review_Report_2007_12_03v3.pdf (Accessed February 27, 2008)

18. Mandel R. A Half Century of Peer Review, 1946-1996 (NIH Division of Research Grants, Bethesda, MD, 1996).

19. Landy, FJ, Farr, JL, Performance rating. *Psychological Bulletin* 1980;87:72-107.

20. Arkes, HR. The nonuse of psychological research at two federal agencies. *Psychological Science* 2003;14:1-6.

21. "Working double-blind." Editorial: *Nature* 2008;451:605.

22. van Rooyen S, Godlee F, Evans S, Smith R, Black N. Effect of blinding and unmasking on the quality of peer review: a randomized trial. *JAMA* 1998;280:234-7.

23. Godlee F, Gale CR, Martyn CN. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial. *JAMA* 1998;280:237-40.

24. Justice AC, Cho MK, Winker MA, Berlin JA, Rennie D, et al. Does masking author identity improve peer review quality? A randomized controlled trial. *JAMA* 1998;280:240-2.

25. Budden AE, Tregenza T, Aarssen LW, Koricheva J, Leimu R, Lortie CJ. Double-blind review favors increased representation of female authors. *Trends Ecol Evol.* 2008;23:4-6.

26. Wenneras C, Wold A. Nepotism and sexism in peer review. *Nature* 1997;387:341

27. Rosenblatt J. Funding innovative research : NIH New Innovator Awards. *ASCB Newsletter*. February 2008; page 2.

28. Bridges to Independence report: NRC. 2005: http://www.nap.edu/catalog.php?record_id=11249#toc (Accessed February 13, 2008)

29. Garrison HH, Stith AL, Gerbi SA. Foreign postdocs: the changing face of biomedical science in the U.S. *FASEB J.* 2003;17:2169-73.

30. Stephan, PE. 2004. What data do labor market researchers needs? A researcher's perspective. In: The U.S. Scientific and Technical Workforce: Improving Data for Decisionmaking, eds. T.K. Kelly et al. Arlington, VA: RAND Corporation.

31. Heinig et al., Sustaining the Engine of U.S. Biomedical Discovery. *NEJM* 2007;357:1042-7.

32. Kaiser J. Med Schools Add Labs Despite Budget Crunch. Science 2007;317:1309-10.

33. Non Tenure Track Faculty: The Landscape at U.S. Institutions of Higher Education, Center for the Education of Women, U. Michigan report (1996): http://www.umich.edu/~cew/PDFs/NTTlandscape06.pdf (Accessed February 14, 2008).

34. Rating of Grant Applications subcommittee of the Committee on Improving Peer Review, Report of the committee. 1996: http://grants.nih.gov/grants/peer/rga.pdf (Accessed February 28, 2008).

35. Mayo, NE, Brophy, J, Goldberg, MS, Klein, MB, Miller, S, Platt, RW, and Ritchie, J, Peering at peer review revealed high degree of chance associated with funding of grant applications. *J. Clin. Epidem*. 2006;59:842-8.

36. Division of Research Grants, Peer Review Trends: Workloads and Actions of DRG Study Sections, 1980-1990. NIH, 1991.

37. Martin M. History of priority scores at the NIH: http://grants.nih.gov/grants/peer/prac/prac_sep_2005/martin_presentation.ppt . (Accessed February 28, 2008).

APPENDIX I:

PREVIOUS AND ONGOING PEER REVIEW EXPERIMENTS

1. Reducing Administrative Burden of Applicants, Reviewers, and NIH Staff

Prebuttals

- From 2004 to 2006, NCI conducted an experiment evaluating the efficacy of prebuttals that enabled reviewers to ask a few critical questions of applicants during the review meeting aimed at clarifying issues about an application before scoring. To address concerns of reviewers, applicants, and NCI program staff about loss of information that would have been gained during site visits, a teleconference with the applicant group was scheduled during each review so the reviewers could ask the applicants up to 4 to 6 "key questions" per project or core. After the teleconference, the reviewers discussed how the answers to the questions affected their ratings for the projects or cores. NCI staff described the outcome of the prebuttal experiment as mostly negative. Reviewers commented that the review discussion was often repetitive, with each issue being discussed three times: before, during and after the teleconference. In most cases, the answers received during the teleconference did not significantly affect the final project or core ratings.

- Other domestic agencies that have incorporated prebuttals in their review procedures include the Burroughs Wellcome Fund and the HHMI Cloister Program. International agencies that include prebuttals in their review procedures include the UK Medical Research Program, and the Australia National Health and Medical Research Council.

Expedited review for early-career investigators

- Since FY 2001, the National Institute on Deafness and Other Communication Disorders (NIDCD) has employed an approach to identify a subset of applications (R01s to early-career investigators) that require only modest revisions to reach a funding threshold. These applications are funded administratively if an applicant were given the opportunity to present corrective measures evaluated by the funding IC and approved by that IC's National Advisory Council. Based on an analysis of the application and summary statement, early-career investigators are invited by NIDCD program staff to submit a 5-page "Letter of Response" for consideration by two members of the IC's council. These Council members lead a discussion about the merits of the response, and those judged to be of sufficient merit are placed into a "high program priority" category and can be supported without the need for an amended application. This practice has led to an increased percentage of early-career investigator R01s being funded by NIDCD.

2. Enhancing the Rating System

Rating criteria

- In 1996, the Rating of Grant Applications subcommittee of the Committee on Improving Peer Review considered issues relating to how reviewers assign scientific merit ratings to applications (34). The Committee felt that the general rating system used by the NIH worked reasonably well. However, with decreasing percentages of applications being funded, the group felt it was important to ensure the highest reliability in scoring, and the greatest amount of useful information for NIH program staff. The Committee made 10 recommendations specifically regarding the rating of grant applications:

  - The proposed, reformulated review criteria should be adopted for unsolicited research project grant applications. The three, reformulated criteria are:

    1. *Significance*--the extent to which the project, if successfully carried out, will make an original and important contribution to biomedical and/or behavioral science
    2. *Approach*--the extent to which the conceptual framework, design (including, as applicable, the selection of appropriate subject populations or animal models), methods, and analyses are properly developed, well-integrated, and appropriate to the aims of the project
    3. *Feasibility*--the likelihood that the proposed work can be accomplished by the investigators, given their documented experience and expertise, past progress, preliminary data, requested and available resource, institutional commitment, and (if appropriate) documented access to special reagents or technologies and adequacy of plans for the recruitment and retention of subjects.

  - Reviews should be conducted criterion by criterion, and the reviewers' written critiques should address each criterion separately.
  - Applications should receive a separate numerical rating on each criterion.
  - Reviewers should not make global ratings of scientific merit.
  - The rating scale should be defined so that larger scale values represent greater degrees of the characteristic being rated, and the smaller values represent smaller degrees.
  - The number of scale positions should be commensurate with the number of discriminations that reviewers can reliably make in the characteristic being rated. An eight-step scale (0-7) was recommended on the basis of the psychometric literature; however, a maximum of 11 steps (0-10) would be acceptable.
  - The rating scale should be anchored only at the ends. The performance of end-anchors should be evaluated and other approaches to anchoring should be investigated as needed.

- Scores should be standardized on each criterion within reviewers and then averaged across reviewers. The exact parameters for this standardization should be defined by an appropriately constituted group.
- Scores should be reported on the eight-point scale used by reviewers in making the original ratings. Scores should be reported with an implied precision commensurate with the information contained in the scores. Two significant digits are recommended.
- If a single score is required that represents overall merit, it should be computed from the three criterion scores using an algorithm that is common to all applications. The Committee favors the arithmetic average of the three scores: however, an appropriately constituted group should test and choose the algorithm to be used.
- Only one of these recommendations was adopted: "Reviews should be conducted criterion by criterion, and the reviewers' written critiques should address each criterion separately."

➢ Since 2006, NINDS has conducted an experiment to ensure that reviewers address all of the special review criteria for applications submitted in response to RFAs or PAs with special review criteria. To accomplish this, NINDS developed a form with a set of questions that addressed each of the special criteria and required the reviewers to submit their critiques on the form. The outcome of this experiment has been generally positive. Given that virtually all of the reviewers addressed all of the special criteria, the resulting discussions permitted more complete evaluation of the applications. The reviewers reported that they greatly appreciated the forms because it helped them focus their critiques. NINDS staff also found that the use of the forms reduced the prose in critiques that simply reiterated what was being done, and did not provide evaluative comments. The use of this special review criteria rating form by NINDS has become routine for some applications.

➢ Several outside agencies rely on separate criteria scores and an overall score including: DOD Congressionally Directed Medical Research Programs, Susan G. Komen Foundation, and HHMI Medical Scholars Program.

Normalization: binning, ranking, percentiling

➢ Binning, or partitioning continuous data into discrete groups, has been previously piloted by NIAID and NIGMS. NIAID piloted a three-bin system in 2004 for Bioshield Initiatives and a five-bin system in 2005 for sites within the Units for HIV/AIDS Clinical Trials Networks. While these scoring systems required additional training and guidance for reviewers, increased flexibility was given to program staff to balance scientific merit and programmatic needs. NIAID felt that binning may be useful for complex applications and initiatives in broad areas of science and may use a binning system in the future.

- From 1995 to 1996, NIGMS piloted a binning system, using the current, initial scoring system (41-point scale), but then rounding averaged priority scores to the nearest 10. This approach maintained the 41-point scale (rather than a 401-point scale), and percentiles were rounded to the nearest odd number to generate a 50-point scale (rather than a 100-point scale). Program staff and applicants in the NIGMS binning pilot had few concerns with this system, and program staff felt they had more flexibility in funding decisions. The system was not adopted, however, due to Council concerns with potential loss of information.

- Normalization and percentiling has been an issue of discussion in the NIH scoring system since 1972 when the practice was first piloted, but then discontinued due to IC objections. Normalization was revisited in 1977 (by the Peer Review Study Team) and in 1979 (by the NIH Committee to Study Priority Scores). Recommendations to use normalized scores again faced IC objections. In 1988, ICs were given the flexibility to use percentiles if they chose to, when the EPMC Group on Movement of Priority Scores recommendation to use percentiles was implemented.

- Ranking as an explicit process has previously been suggested as another means of standardizing scores. A 2006 study reported high variability in funding decisions when based on two reviewers, and instead recommended 10 reviewers using a ranking method to provide optimal consistency (35). The Burroughs Wellcome Fund and HHMI Training Awards both use a combination of binning and ranking. The Australia National Health and Medical Research Council uses a final ranking system at the end of review sessions. However, it is interesting to note that re-ranking applications at the end of study section was determined to have minimal impact on final outcome in a 1991 DRG study (36,37).

- The UK Medical Research Program uses a 10-point scoring scale; NSF uses a five-bin system; and the Burroughs Wellcome Fund and HHMI Training Awards both use a combination of binning and ranking (Burroughs Wellcome: binning with top 10 rank ordered, HHMI Training Award: binning used in first round, scores and rank order in second round).

Shortened applications

- Several ICs have explored the use of a shortened application. The National Heart, Lung, and Blood Institute (NHLBI) began a pilot in early 2008, in which the length of the Research Plan (items 2-5 of the PHS 398 form) was shortened. On the basis of one review round thus far, it appears that the applications contained sufficient critical information for review, and applicants did not use appendices to provide additional materials. However, reviewers reported that the shortened application did not reduce their burden, suggesting that even shorter applications may be necessary to reduce reviewer burden. NIAID has used 15-page applications (with no appendices permitted) for Project Bioshield initiatives. NIAID staff reported that these shorter applications did not appear to affect the

ability of review panels to assess scientific merit, although a caveat is that these applications are, by nature, very focused.

- Under the leadership of the National Institute of General Medical Sciences (NIGMS), the NIH Roadmap has made use of shortened applications for both the Pioneer Award (5 pages, reviewers each review ~20 applications) and the New Innovator Awards (10 pages, reviewers each review ~35 applications). For the New Innovator award, an applicant must emphasize the significance of his or her project; what makes the project exceptionally innovative; and, his or her qualifications. No preliminary data is required. While it is not yet possible to assess the scientific contributions of awardees chosen in this manner, reviewers appear to be satisfied with the process-level aspects of this approach.

3. Enhancing Review and Reviewer Quality

Editorial board review

- NIAID conducted an experiment with a two-stage editorial board style review, the first tier rating technical merit, and the second tier rating global impact toward a particular field or health condition. The first-tier technical review was held via teleconference. One application per teleconference was reviewed with approximately 20 non-conflicted panel members. A minimum of 2 non-conflicted members from this review panel were also members on another review panel to ensure consistency across the review panels. Members from each of the review panels (as well as experts required to review the other RFA requirements) convened at a face-to-face "Overall Merit and Impact" review. Draft summary statements were provided to this second panel, prior to the meeting. Positive outcomes of the experiment included that the first-tier review ensured appropriate evaluation of the scientific and technical merit of proposed projects using a teleconference platform that facilitated the participation of more senior national and international reviewers. Negative aspects were that this type of review is time- and resource- intensive, and requires a large number of staff.

Electronic-assisted review

- CSR is conducting an experiment with Asynchronous Electronic Discussion (AED) Review. This new method, based upon the use of a threaded message board with features tailored to NIH review, permits the asynchronous discussion and private scoring of grant applications without the need for concurrent assembly or teleconference.

- CSR is also evaluating the usefulness of Video-Enhanced Discussion (VED), with the long-term goal of expanding the use of this technology to over 100 review meetings per year to provide a virtual face-to-face meeting environment. The VED approach is an option for SROs to conduct real-time, virtual face-to-face meetings with a sizable number of applications reviewed simultaneously.

- The National Institute of Mental Health's (NIMH) experience with AED has been mostly negative due to the labor-intensity of the process. Although refinement of the implementation of AED is likely to reduce some of these shortcomings, NIMH remains cautious about the wide spread implementation of AED as a review platform.

- In 2007, NIAID conducted a peer review experiment using an Asynchronous Electronic Discussion (AED) with an Internet-Assisted Review (IAR) scoring process for approximately 225 Loan Repayment Program applications. Reviewers reported generally preferring the flexibility of this method over teleconferencing. NIAID now conducts all LRP reviews this way as a result of this experiment.

- NIMH staff have found that VED is more easily implemented, and NIMH has employed this review platform for both small review meetings and conferences. Reviewer response has been very positive.

- NIMH has also conducted several review meetings involving 15 to 30 applications using teleconference with Web-based White Board (Adobe Connect). The latter allows for chat, video, white board, and file sharing. NIMH's experience with this approach has been highly favorable, and feedback from reviewers also indicates a high degree of satisfaction. NIMH will continue to experiment with VED and Adobe Connect as an adjunct to teleconferences.

- In 2005, 2006, and 2007, NCI conducted experiments with electronic review for its NIH Loan Repayment Program (LRP), which requires a large number of reviewers. NCI used a fully electronic/virtual review using the NIH Internet-IAR System. Three reviewers were assigned to each application and each reviewer was assigned eight applications. Outcomes were better than expected and were quite satisfactory. The entire review process, which is otherwise equivalent to four or more standard review meetings, can be executed by a fully dedicated SRO and a single, competent support staff person.

- In 2006, NIAMS conducted an experiment to evaluate the efficacy of IAR for reviewing RO3s. Pilots were completed for 3 to 4 rounds with no complaints from applicants. Reviewers reported becoming increasingly comfortable with the process, and some have suggested that a chat room or teleconference would further facilitate the discussion during review. In addition, NIAMS found that the use of IAR made it possible to recruit more highly qualified reviewers, who would otherwise be unable to participate in the review.

- In 2000, NIAID initiated the development and implementation of a secure electronic review (ER) website to capture reviewer critiques for grant applications and contract proposals. NIAID staff piloted the use of the ER site, and other ICs used this site for several years prior to the development of the NIH-based IAR.

- In 2004, NIAID developed the "The Reviewer Support Site (RSS)," a secure, Web-based system that enables reviewers to access review meeting documents via the Internet. A newer version, RSS Version 2 is now being used widely by NIAID and will soon be available for use by all ICs. The new system offers reviewers a single URL to handle all aspects of review from completing conflict of interest forms to posting critiques. NIAID staff and reviewers relate positive feedback about its usefulness.

- Outside agencies that have incorporated electronic review components in their peer review practices include:

    - DOD Congressionally Directed Medical Research Programs
    - NSF
    - Susan G. Komen Foundation
    - Burroughs Wellcome Fund
    - Canadian Institutes of Health Research

Reviewer training

- In 2005, NIAID developed an SRO training program and reports favorable results as evidenced by its ongoing expansion and use by other ICs. The program contains 28 training modules and is aimed at standardizing the implementation of review policy, procedures, and practices, and Web site (http://ai-appewp1.niaid.nih.gov/srp/).

- Several ICs have conducted experiments of reviewer orientation teleconferences to promote the consistent application of review criteria, particularly for complex applications. Favorable outcomes include: i) reviewers generally report improved ability to focus critiques on issues most germane to program announcements; ii) time spent discussing review and programmatic issues at review meetings is often reduced; iii) panel members report valuing the opportunity to discuss review and programmatic issues with NIH staff; and iv) conference calls often enhance interactions between NIH review and program staff. Several ICs reported that the benefits of orientation teleconferences outweighed the costs.

Reviewer recruitment

- Since 2003, the National Institute of Neurological Disorders and Stroke (NINDS) has conducted an experiment to address the competition between study sections for a sometimes very limited supply of senior reviewers with specific technical expertise in emerging areas. To enable more than one committee to share such reviewers without asking them to make two trips to the Washington, DC area, the committees have scheduled their meetings for consecutive days in the same hotel. NINDS standing committees now routinely schedule their meetings on consecutive (or overlapping dates in cases of multi-day meetings) so that if there is a need to share reviewers, the option will be available.

- Since 2004, NIMH has conducted experiments to evaluate the use of public reviewers as full voting members on NIMH review committees for treatment and services research. Public reviewers are asked to comment on specific aspects of grant applications including: feasibility of the study, the potential for public health impact, human subject protections, and recruitment of an appropriately diverse sample. Individuals interested in serving as public reviewers typically are mental health consumers or their family members, mental health professionals, members of advocacy groups, educators, etc. Public reviewers participate in the review discussion, write critiques, and provide scores. NIMH has found that the input provided by public participants adds meaningful information and sensitivity to the review process. Several outside agencies incorporate community advocates in their review session panels.

4. Optimizing Support at Different Career Stages

Retrospective awards

- NCI had an unsuccessful career recognition award, which consolidated all the awardees funding into one stable stream for an extended period of time. However, this award had negative impact on the awardees' productivity, and NCI has discontinued the program.

- A similar, highly successful program is the Howard Hughes Medical Institute (HHMI) Investigators program. Since the early 1990s, HHMI selected investigators through rigorous national competitions with the aim of identifying researchers with the potential to make significant contributions to science. HHMI is guided by the principle of "people, not projects." Investigators are appointed for an initial term of 5 years, and upon successful scientific review are renewed for additional terms of 5 years. Renewal of an investigator's 5-year appointment is dependent on a peer-review process that is both retrospective, evaluating the originality and creativity of the investigator's work relative to others in the field, and prospective, evaluating the investigator's plan for future research.

6. Reducing Stress on the Support System of Science

Minimum percent effort

- HHMI investigators are expected to devote at least 75 percent of their total effort to the direct conduct of biomedical research during the period of HHMI support (5 years, with possible renewal). These investigators may spend up to 25 percent of their effort on related activities, such as teaching, consulting, and administrative duties. HHMI pays the entire salary of the principal investigator, and while it encourages its investigators to seek competitive external research grant support, commercially sponsored research--industry funding under terms that would give the funder rights to intellectual property developed in an HHMI laboratory--is not permitted.

APPENDIX II:

TRANSCRIPT OF FEBRUARY, 21, 2008 ACD TELECONFERENCE:

COMMENTS ON INTERIM DRAFT REPORT

The official transcript of the ACD teleconference discussion is posted online at http://enhancing-peer-review.nih.gov/ .