NIH Regional Consultation Meeting on Peer Review – San Francisco
October 25, 2007
Meeting Summary

Meeting Context and Review of Ongoing Activities

Dr. Lawrence Tabak
Director, National Institute of Dental and Craniofacial Research, NIH; Co-Chair of the Working Group of the Advisory Committee to the NIH Director (ACD) on NIH Peer Review

We are doing this study in partnership with the scientific community to strengthen peer review in these changing times. The breadth, complexity, and interdisciplinary nature of science have increased, and that's good. But it also creates challenges to the system we use to support biomedical and behavioral research. I emphasize “system,” because the charge of our group is to go beyond peer review, per se. NIH must continue to adapt to fields of science that are rapidly changing and to ever-growing public health problems. Whatever we do, we must ensure that the processes are both efficient and effective for applicants and reviewers alike.

Two committees were created to help oversee this program. The first is the Working Group of the Advisory Committee to the Director (ACD) of NIH, which Dr. Yamamoto and I co-chair. A number of people on this committee have served on the previous formal report of peer review, the so-called “Boundaries Report.” The second, equally strong group of individuals comprise the internal NIH Steering Committee Working Group on Peer Review, which I co-chair with Dr. Jeremy Berg, director of the National Institute of General Medical Sciences.

The Center for Scientific Review (CSR) is looking at some of the mechanical attributes of peer review. Dr. Toni Scarpa and Dr. Norka Ruiz Bravo are on these committees, so we can assure you we are coordinating our efforts with the CSR initiatives.

We began our study with the soon-to-be-completed diagnostic phase. We've had a request for information (RFI), where we asked stakeholders for feedback in six specific areas: challenges to the NIH support system, specific challenges related to the NIH peer review process, solutions to these challenges, core values of the peer review process, the criteria we use to evaluate grant applications and the scoring system we use, and whether the current process is appropriate for investigators at different stages of the career pipeline.

Although the timeline for the RFI has ended, you can still send input to: peerreviewrfi@mail.nih.gov.

Additional activities include the following:

- Dr. Yamamoto and I held two teleconferences with about 100 deans or their representatives.

- We held a series of meetings like this one around the country.

- We asked members of the ACD Working Group, deans, and NIH institute directors to select individuals to serve as liaisons to further broaden outreach to the scientific community.

- We created a common Web site.

- The internal Steering Committee collected input from the institutes and centers. We held three meetings like this one internally at NIH.

- We reviewed the very robust literature about peer review.

- We are reviewing how other agencies approach peer review, both in the United States and around the world.

- The National Science Foundation released a report in August on the impact of proposal and award management mechanisms, which also will inform our efforts.

NIH leadership soon will begin to determine next steps, and these likely will include pilot experiments. The design and initiation of these pilots, and their associated evaluations, should be done by late winter or early spring. This will lead to the development of an implementation plan, which we will disseminate widely. The pilots that are successful will be expanded, and that will lead to the development of a new peer review policy.

Goals for the Meeting

Dr. Keith Yamamoto

Executive Vice Dean, School of Medicine,UCSF; Professor, Cellular/Molecular Pharmacology and Biochemistry/Biophysics,UCSF; Co-Chair of the Working Group of the Advisory Committee to the NIH Director (ACD) on NIH Peer Review

The title of my presentation comes from the charge that Dr. Zerhouni gave the Working Group: “Fund the best science, by the best scientists, with the least administrative burden.” That breadth of charge is both exhilarating and daunting.

Let me step back a long way to tell you a quote from the surgeon general in December 1945 – literally the month before the current NIH peer review system was adopted: "The only possible source for adequate support of our medical schools and medical research is the taxing power of the federal government. . . Such a program *must assure complete freedom for the institutions and the individual scientists in developing and conducting their research work.*"

Let's think about where we are now. Tom Cech, Nobel Laureate and head of the Howard Hughes Medical Institute, said in 2005: "Discovery and innovation are to some extent taking place in spite of, rather than because of, the current policies and practices of major biomedical funding agencies.”

Those two quotes cover a fair amount of distance, and the challenge before us today is to ask what's happened that's taken us to this point, and what can we do to try to recover some of that initial ideal.

Intrinsic conflicts are built into the whole concept of peer review. One is reviewer self-interest; this acknowledges that reviewers are aware that their judgments affect funds from the same pot of resources they compete for themselves. The second conflict is reviewer conservatism. To have the best system, we need to have the best scientists participate in the review process. These scientists created the prevailing paradigms, and so they will defend them. They will also favor ideas from people who think the same way they do.

We also need to respond to the fact that the doing and reviewing of science are continuously changing in dramatic ways. Investigators will be able to craft proposals with a far broader scope than before. Technology continues to be a major driver in research, much more than in times past. This creates more complexity, requiring more expertise on the part of the investigator and more people to get the job done, often in multidisciplinary teams.

You probably hear colleagues say, "If there were just more money in the system, there wouldn’t be a problem." I think if there were money in the system, there would be fewer people at this meeting, but the issues would still be there because of those changes. We need to think about ways to create a system for supporting science that keeps pace with change.

Twenty years ago, NIH used 1,800 reviewers; last year, it used 18,000. Some 80,000 grant applications came to the NIH – a doubling over the last few years. This resulted from the doubling of the NIH budget between 1999 and 2003, followed by its flattening – a decline of 13% in the spending power of the NIH dollars over the last 4 years. During the doubling, we established new institutes, hired new faculty, and built new buildings. When the budget flattened, we suddenly had a big pool of investigators scrambling for money, so the number of applications exploded, and these were much broader in scope than before. Study sections expanded to dozens of members instead of the15 to 20 mandated by Congress, and almost all were ad hoc.

The study section culture also changed. With fewer senior scientists to provide wisdom and perspective, and in part because of the perception of limited resources, the process took on an adversarial tone rather than the supportive, collegial tone of earlier times.

The peer review system clearly needs to evolve and adapt to these changes. We are looking for bold thinking in all areas. The following, and more, are open for discussion:

- Reviewer Criteria and Focus: Are there ways to reestablish the importance of the idea, of the impact of the work were it to be done well? NIH has always supported projects; should it be paying more attention to supporting people? Should there be two tracks: one to support people, and one to support projects?

- Application Structure and Content: Do NIH grants need to be the longest in the world? Are there ways to deemphasize preliminary data and methods by shortening the grant applications and refocusing and changing what's mandated in the content?

- Reviewer Mechanisms and Mechanics: Does the study section system as it was crafted really work with today's grant applications of enormous scope? Are there other ways to pull together needed expertise? The editorial board model refers to an idea that the 15 or 20 study section members (similar to an associate editorial board at a journal) would, upon receiving a grant application, send it to reviewers around the country and ask for focused comments based on their particular expertise. That input would give the study section some confidence to look at the larger picture. How important are face-to-face meetings?

- Reviewers and Review Culture: Are we using reviewer time and expertise effectively? Are there ways to regain the kind of value of service that motivated people to participate in the process, and that made them feel it was an honor, a duty, and something worthwhile they actually enjoyed?

- Scoring: NIH uses a hard numeric scoring system that goes out to two decimal places. Should we think about moving to ranking rather than scoring?

We cannot do these things piecemeal. It requires broad changes that have a broad impact, simultaneously addressing multiple things. We need to be rigorous but fair, and we need to:

- Acknowledge the special needs of both new and established investigators.

- Ensure that the process both celebrates and supports innovative research.

- Ensure that the process is respectful of the applicant and the reviewer, and is recognized as something that addresses the needs of both groups.

Statements/Proposals from Consultation Groups Offering Specific Strategies or Tactics for Enhancing NIH Peer Review and Research Support

D.L. Jewett, M.D., D.Phil., Professor Emeritus, University of California at San Francisco; Research Director, Abratech Corp.

I've watched this process evolve and have seen that the reviewers have taken control. They ask each other, or the SRA, what scores will get funded, and then they score the ones they want to fund.

I've seen at least four NIH directors come through and try different initiatives to get more translational research going. But when I try to put in a basic science proposal that will have clinical applications, the study sections do not like it, and they vote it down.

Feedback in the system has been lacking for several decades. I propose that PIs be able to submit 1-page comments about their critiques. If there are major difficulties, these can be identified so some action can be taken other than just rejecting the proposal. This gives the PI the opportunity to point out misunderstandings and mistakes in the review, and to reinforce correct judgments/opinions.

The PI comments would provide information regarding the quality of the criticisms of the reviewers. This could help the chair and SRA evaluate the performance of reviewers and the study section as a whole. The comments could also flag inadequate reviews that may need to be handled in another way, such as by obtaining additional reviews (as may be possible if there is a shift to utilize more written reviews from non-members of the study section). Any additional reviews should also be available to the PI for comments.

The PI comments could be categorized and used to evaluate CSR as it evolves under various initiatives and programmatic changes. The comments could also be used as part of an evaluation of the performance of the SRA and study section chair. In order to increase the ability of the higher levels of CSR to improve quality (by early detection of poor performance), the PI should be offered the opportunity to "bcc" the comments to a CSR level above the SRA. Of course, a single comment will have little weight, but numbers of such "red flags" from independent PIs could be important in indicating problem areas. Finally, CSR can report to the Office of the Director regarding the statistics of categorized comments, thus providing OD with assurance that PI feedback is part of the evaluation of CSR performance. The lack of such feedback in an institution the size of NIH is an obvious flaw in the design and organization of peer review.

I recommend that reviewers number the lines on their reviews so it will be easy to refer back to them.

Finally, preliminary data are essential because the study sections require success. I've always found that any suggestion past preliminary data is always rejected as too speculative.

Peter Bacchetti, Ph.D., Professor of Biostatistics, University of California at San Francisco

My proposal is to discourage consideration of sample size in reviews of grant applications and to eliminate sample size justification from RFA and RFP requirements. I contend that reviewing sample size provides no benefit but is, in fact, harmful to the peer review process and to science. [See Attachment A for handout that complements this presentation.]

The usual convention is to calculate a sample size that will produce at least 80% power. A myth has grown around this convention that any study with less than 80% power is doomed to be worthless, and any study with 80% power or more will be definitive. In reality, the scientific or clinical value that a study can be projected to produce increases gradually as the proposed sample size increases. Nothing special happens at 80% power.

Furthermore, the projected value of a study exhibits diminishing marginal returns as sample size increases. Reference #2 establishes this in detail and argues that reviewers therefore do not need to verify that a study will have at least 80% power.

An additional problem is that the power of a study cannot be verified in advance. For any interesting study, the inputs needed for the calculations will be uncertain, and small changes in the inputs are magnified in the resulting sample sizes. The inaccuracy of sample size assumptions is theoretically inevitable and has been empirically verified.

The inherent murkiness of sample size calculation makes it easy to show 80% power for any study. And because every proposal claims to have at least 80% power, such claims provide no information about how valuable any particular study is likely to be.

Review of sample size harms the peer review process by encouraging dishonesty in both investigators and reviewers. Investigators must often base sample size on cost and other practical limitations, but they regard admitting this as peer review suicide. So they ignore the actual sample size reasoning when writing grant proposals, and instead make a conventional case for at least 80% power. Any reviewer who is so inclined can question the assumptions of any sample size calculation. This means that whether a sample size justification passes peer review is up to the discretion of the reviewer, a bad situation in a process that seeks to be fair. Because criticism of sample size plans is always possible and very common, it is an ideal mechanism for facilitating discrimination and cronyism.

Current sample size conventions harm science in two ways. First, they contribute to the widespread fallacy that $p>0.05$ can be interpreted as proving there is no difference or no effect. Many researchers, peer reviewers, and editors implicitly accept the myth that a study will be definitive if it was calculated to have at least 80% power, so when a study is not definitively positive, it is interpreted as definitively negative, even when point estimates suggest a positive interpretation. [Attachment A, #7, provides an example.] The second source of scientific harm is bias against innovative proposals. Innovators struggle to somehow prove that their studies will have 80% power, but doing so in any meaningful and honest way is impossible for innovative studies.

Bonnie Blomberg, Ph.D., Professor, Department of Microbiology and Immunology, University of Miami Miller School of Medicine; The American Association of Immunologists (AAI)

AAI shares NIH's goal to maintain and enhance the quality of the review process. Following are AAI's positions on some changes being proposed, or already implemented, by NIH:

.

- Don't shorten review meetings to 1-day face-to-face time. This provides less time to discuss each application and diminishes the key incentive for reviewer service.

- Don't increase the number of phone reviewers or the participation of reviewers by posting comments on electronic bulletin boards. Face-to-face meetings are the gold standard that safeguards the quality of the review process. They are more efficient and effective and enable valuable informal discussions during breaks.

- Don't allow the mandatory triage of greater than or equal to 50% of applications simply by averaging the scores; this gives too much power to individual reviewers. Not doing this would help to ensure fairness and the perception of fairness. In addition, all review group members must be able to request discussion of any application.

- Don't use inexperienced junior scientists as reviewers. PIs should only serve several years after they have reviewed their own NIH grants, and only as ad hoc reviewers until they have achieved the rank of associate professor or equivalent. That protects them and us.

- Don't adopt a two-stage review process. The experience of the study section includes the ability to discriminate significance. Two-stage review would duplicate effort, slow the process, and discourage participation by experienced PIs who would find providing this limited review less rewarding. We have already done this experiment with DOD.

- Don't fund investigators rather than projects, or projects without regard to the investigator. An outstanding project still requires an outstanding investigator. Prominent investigators do not guarantee quality projects, and new investigators deserve a full and fair chance to succeed.

Following are AAI's recommendations:

- Allow regular review group members to serve only 2 times a year, to ensure they get a break in the process.

- Provide reviewers with an additional 4 to 6 weeks rather than the current 14 days to submit their own applications during cycles in which they serve.

- Continuous submission might work very well, but concerns about the use of special emphasis panels (SEPs) – including frequent member turnover and the fact that revised grants are sometimes assigned to entirely new SEPs – require that this idea be thoroughly discussed and evaluated before a new system is implemented.

- Ensure a fair policy governing the assignment of study section members' grant applications, including allowing members to have input into the appropriate assignment of their applications.

- Provide all unscored applicants with the quartile in which their application was ranked to help them decide whether to revise and resubmit, or start anew.

- Review scores and ranking of applications at the end of study sections. Relative scoring behavior frequently shifts during the course of the review meeting, resulting in inequities in scoring based upon whether the grant was discussed early or late in the meeting. AAI supports the initiation of a process to revisit grant scores before the end of the meeting, and possibly rank them relative to one another, as is currently done in NSF reviews.

- Don't significantly increase the number of grants per reviewer, even if the length of R01 applications is decreased. The amount of time required for a well-informed and thoughtful review is not directly proportional to application length.

- Conduct and publicly disseminate objective, thorough, and transparent evaluations of all pilot programs undertaken by CSR and NIH.

- Shorten the length of the application from 25 to 15 pages, but (1) allow the inclusion of an optional 5-page supplement by new PIs to provide additional evidence of technical feasibility and other details necessary to ensure a fair review; and (2) the limit should not include the 3-page introduction to resubmitted grants, since additional space is often needed to respond to a previous review.

- We support the recent change requiring reviewers to consider the significance of proposed research.

William M. Burdon, Ph.D., University of California at Los Angeles, Integrated Substance Abuse Programs (ISAP)

UCLA ISAP investigators are conducting research under 25 projects funded by NIH. In the current budget year, these projects represent almost $8 million in direct expenses.

Following are our recommendations:

1. Develop and implement an online application review system. This process would greatly improve the reliability and consistency of application scoring, and significantly improve the quality and breadth of proposed studies.

    - Combined with the recently implemented online grant submission system, which verifies compliance of submitted applications with established NIH guidelines, an online application review system would allow grant reviews to be conducted much sooner following submission, perhaps as soon as 4 to 6 weeks.

    - With an online grant review system, applications could be reviewed by individuals selected from an expanded pool of qualified application reviewers. NIH would develop this pool by establishing qualifying criteria and inviting a broad range of individuals from around the country to apply online.

- Registered grant reviewers would be assigned to virtual review committees formed separately for each funding cycle, or for the review of applications submitted under RFAs. Within each of these committees, reviewers would be classified in terms of their areas of expertise, which would determine the areas they would be asked to critique and score. The combined scores from these reviewers would make up the full application score.

2. Develop a process for reviewing resubmitted applications. Implicit in current NIH guidelines and requirements for resubmitted applications is the understanding that improvement in an application's priority score and its subsequent chance of being funded is dependent upon the extent to which the revised application adequately addresses the concerns of previous reviewers. If reviewers are permitted to critique and score resubmitted proposals as completely new applications without deference to the original reviewers’ critiques or the changes made by the investigator in response to those critiques, the ability of the investigator to improve a proposal to a potentially fundable threshold within three attempts is undermined. To alleviate, if not eliminate, this flaw, we recommend that application scores below a certain threshold (e.g., priority score of 1.4 or less, or a 20th percentile ranking) be classified as potentially fundable applications that investigators can resubmit. Such resubmissions would be reviewed and scored on the basis of how the applicant responded to and addressed previous reviewers' comments.

3. The current presumed 5-year maximum for conducting studies is obsolete. The length of proposed study should be open-ended, or at least extended significantly. For funded grants that extend beyond 5 years, NIH institutes could conduct periodic progress reviews after the fifth year to determine whether to continue funding.

4. One of the primary distinguishing features of the various funding mechanisms is the period covered and the amount of funding available. We believe that competently designed research should determine the length and required funding of proposed studies. The mechanism structure should be limited to organizing types of applications (e.g., exploratory studies, clinical trials, etc.). With some limits, the length of a proposed study and the amount of funding requested can become part of the application and the review process.

David Busath, M.D., Professor, Department of Physiology and Developmental Biology, Brigham Young University

The current NIH approach is successful because it is based on competition and a free market, but it is anticipatory rather than product-oriented, and the judgment is peer-expert-based, which can be frustrating and possibly inefficient.

I propose that NIH develop a complementary avenue to fund products, i.e., papers, in three stages: draft, publication, and citation. It could be readily implemented, mostly automated, sufficient to maintain and stimulate the intellectual infrastructure, and affordable.

As the table entitled “Traditional Research Career Trajectory” shows, research careers blossom over a 30- to 40-year time span. Funding on a by-paper basis could easily match the program development costs, as shown in the table entitled “Proposed Standard Reward Trajectory.” [See Attachment B.1. for handout that complements this presentation.]

Strategically, the program would consist of a tiered reward system based on online submission of 1-page “pub-proposals” or publication drafts, as well as funding at the time of publication in a peer-reviewed journal and at the time of published citations [see Attachment B.2 for handout that contains proposed strategic and tactical approaches.] There could be an online review of the pub-proposals, with a low threshold for funding (perhaps 90% getting funded) to weed out the worst work. The funding for publications and citations could be partly merit-based, and it could be automated using impact factors and easily recognized details about the nature of citations.

The disadvantages of such a program might include proliferation of poor-quality publications and citations; growth of shortsightedness and self-contentedness; reduction of peer-review standards by journal reviewers; and failure to address immediate or significant, but non-prestigious, human needs. However, given the solid development of scientific standards for publication, the prestige and discovery-based motivation of scientists, the human nature to serve the public good and to develop productive research programs, and the opportunity to continue stimulating growth in specific areas from the top down with specialized RFAs, and perhaps special recognition of important papers, these disadvantages can be largely overcome.

Ken Dill, Ph.D., Professor, University of California at San Francisco

I propose something very simple that many of us in our home institutions do all the time: ranking proposals rather than scoring them. For example, suppose I have 10 grants to review, and then I give 10 points to my best, 9 points to my next best, and so on. This solves the problem of calibration and normalizing one reviewer to another, and one pile of grants to another, because each reviewer has his or her own internal reference frame. It extracts a lot more information from a given reviewer throughout the review cycle. [See Attachment C for an example of a simple ranking system.]   For example, if I have n grants to read, I'm giving n^2 bits of information by doing this rank ordering pairwise. I'm only giving n bits of information if I do scoring.

We're often asked by our SRAs to spread out our scores. The problem is that natural processes of voting, scoring, and statistics lead to Gaussian distribution functions. There's no way to fix the Gaussian by just telling people, "Please try to flatten it," because the reason we're getting a Gaussian is the collective property of the whole group. So I can't tell any reviewer, "Here's how you can flatten the distribution function," because there's nothing they can do.

Ranking does this sort of thing much more automatically. Ranking also solves some pathologies in our current system. For example, I can walk into a study section and spend 2 days there, and my entire time can be completely useless if I do certain things. If I always vote the average score, then I've contributed no further information for discriminating anything. If I always vote with one of the lead primaries, I also contribute nothing. But with ranking, you contribute a lot of information because you've already done that in advance. NIH calls on you to use your specialized expertise to be able to discriminate, which ranking helps us do a lot better.

Ranking removes the need for face-to-face reviews, which would save NIH money and save reviewers' time. One reason we sit together in study section is to get normalized. If I'm reading one pile of grants, and you're reading another, or I'm a Pollyanna reviewer with a high average score, and you're a cranky reviewer with a low average score, we've got to normalize all of this. Ranking would allow for more independent reviews. A new book called *The Wisdom of Crowds* says groups can sometimes be smarter than individuals in predicting things when three conditions are met: The individuals must be independent, they must have a diversity of opinions, and they must draw on their own specialized knowledge.

Ranking supports advocacy-based review rather than fault finding, because ranking is more or less like voting or making a purchase. You pick something you like as opposed to criticizing something you don't like.

Nejat Düzgünes, Ph.D., Department of Microbiology, University of the Pacific Arthur A. Dugoni School of Dentistry, San Francisco

In an article entitled "Science by Consensus" published in *The Scientist* 8 years ago, I stated that one of the foremost concerns of biomedical scientists in the United States is the difficulty in obtaining grant funding from NIH. I indicated that review panels expect so much preliminary data that the major part of a discovery needs to have been already made, indicating that NIH is not funding actual discoveries, but merely their further characterization. I also pointed out that the tedious description of what a scientist is going to do 5 years from now is an unrealistic exercise in bureaucracy and is contrary to the true nature of scientific research.

The multiple resubmission of grant applications overwhelms the NIH review system and takes scientists away from their research. This crisis cannot be resolved by attempting to incrementally improve a broken system. The solution is a paradigm shift in the way we fund research so that biomedical scientists can stop playing grantsmanship games and focus on their research:

- The majority of NIH funding would be allocated to researchers with a track record of solid publications as determined by large international panels of established and younger scientists.

- The panel members would not travel, and the review process would not require tedious analysis of long proposals, but a simple scoring of accomplishments that have already been peer reviewed for publication.

- The grants would be limited to $300,000 per year for 10-year periods. The awardees would not be permitted to submit R01 applications unless they forfeit their grant within a year, placing them on an equal footing with other scientists vying for R01 support. They could, however, apply for shared instrument and nonfederal grants.

- Young scientists starting out in an independent position and with a proven post-doctoral track record would receive $50,000 a year for 5 years to establish their own research programs.

Benefits of this new paradigm include:

- Awardees would save a lot of time by not having to submit numerous grant applications, avoid the anxiety of uncertainty, focus on their research, take risks, and be able to be creative.

- NIH would save large sums of money and time on the review process.

- Current grant reviewers would save time to spend on their own research, and would save the world tons of carbon footprints by not having to travel to NIH.

- This steady source of funding to young investigators would facilitate productivity and help them obtain supplemental funding from other sources.

- Even if 10,000 new grantees would be added each year, the cost during the fourth year would be only $2.6 billion, less than 10% of the NIH budget.

- The system would distribute scarce resources more efficiently and equitably to scientists with demonstrated merit.

Within the first year of the program, 10,000 such grants would be funded at a cost of $3 billion. The number of awardees would be ramped to 40,000 in 4 years. If 2,000 new long-term awardees would be added each year, the direct costs would increase by only $0.6 billion a year. This is more than double the number of new R01s awarded in 2005. NIH would then focus on evaluating large R01 and program grants. Indirect costs would be limited to 30%, bringing the cost for the 40,000 grants during the fourth year to $15.6 billion, slightly more than half the current NIH budget.

After 10 years, the productivity of scientists in the new funding system – measured by publications, citations, patents, and treatments developed – would be compared with that of investigators who choose to struggle through the traditional process. Regardless of the outcome of the comparison, the new system would be a more humane and rational way of funding biomedical research.

Steve Petersen, Ph.D., McDonnell Chair for Cognitive Neuroscience, Washington University, St. Louis

The low predictability of funding success has produced a grant system flooded with multiple proposals through multiple cycles from each PI. The large number of proposals spawns an equally large number of reviews, overburdening everyone in the system. New reviewers are particularly poorly served by the current structure, now getting their first R0I or independent grant on average when they're in their 40s. This ripples backward into post-docs. And as most of you at big graduate programs know, it's affecting applications to biomedical programs in the United States.

My proposed solutions are based on some simple ideas: (1) It's easier to judge scientists and their past productivity than the specifics of proposals; (2) funding should be relatively smooth and predictable rather than all or nothing and capricious; and (3) more scientists should have access to the grant system.

I propose the following:

- Judgment on proposals should be more retrospective than prospective. For new investigators, at least 65% and more like 75% of funding decisions should be based on training history, available mentoring, and early productivity. For established investigators, 75% should be based on productivity and influence over preceding cycles, and the rest on proposed research. This seems radical, except NIH now gives out more than 10% of its funding under just such a system to its intramural researchers. I would much rather be judged in the intramural system at this point. Further, since a productivity review could be assessed relatively objectively based on things like papers and citations, researchers could make reasonable predictions about their future funding prospects.

- Funding should come in $150,000 modules. Changes to a researcher's funding should be based on review, which, again, should be based mostly on past productivity. Only in exceptional cases of prolonged inactivity or a complete lack of activity would they be terminated from the system when they're already receiving multiple modules.

- New investigators would be reviewed by completely separate committees, and they would be given one module for 3 years, funded at very high rates. They would go through the same system for first renewals and have relatively high rates of renewal and potential incrementing of their funding.

- Except in exceptional cases, no investigator should receive more than four modules, either in single investigator grants or as parts of programs projects. This would encourage the dispersal of funds to a larger number of scientists, including new investigators.

- Investigators could add additional funds through infrastructure, core, and targeted initiatives. Rather than being self-sustaining, these would have to undergo the same sort of review of increased productivity through their past cycles.

Robert Reinhard, Community Advisory Group, San Francisco Department of Public Health

As a community representative, I work with investigators and the networks that fund global HIV/AIDS prevention and treatment trials, where the expertise and input of public representatives is very sophisticated. They are a natural component of the peer review system as a matter of scientific merit and an example of what I think NIH just explained recently in its Partners in Research Program announcement.

There is another way in which these public values can be folded into the peer review system, and that is to harness the scoring so that meritorious projects lead more directly to public health improvements. Following are a few suggestions of how that can be accomplished:

- When applicants are scored, elevate the relative weight given to clinical translation principles so that applicants are asked to document more fully the ways to bring results to the bedside. This applies throughout all phases of research. It would account for contingencies such as approvals, so-called negative results, obstacles, and what the Institute of Medicine has called the T2 block, or second translational block. It would also integrate the roadmap values to feature effective community engagement and interaction with the participants.

- Invigorate the currently archived policy of including public representatives on study sections. The assumption that it's not necessary or appropriate to include public representatives is not supportable today because the teams benefit from public participation. Input from public representatives should stand up to the same rigorous scientific standards as input from any other team member.

- The scoring system would be revised to place diversification of study participants and inclusion of underrepresented populations into a primary score element, and to measure the real effectiveness of those efforts. Right now, the scoring system only looks at the adequacy of a proposal. Representation could be wider than is currently described in current NIH policy, and the applicants would be expected to explain efforts to correct research disparities and to lay out their tracking history. That could include line items in budgets for targeted recruitment.

I know NIH has received a lot of comments about reducing the number of grant categories and funding mechanisms, which is a great efficiency. But when doing so, please ensure that the results do not give short shrift to the science agendas and fiscal year research work plans that the institutes have targeted for study.

Fred Schaufele, Ph.D., Professor, University of California at San Francisco

During the course of a study section, we have to find ways to take into account the ebb and flow. What Ken Dill mentioned about doing rankings beforehand really would help to mitigate those problems, so I would strongly support ranking as your primary scoring mechanism.

One of the problems I see in the study section is that we often don't know exactly how to judge relative significance or innovation in relation to the investigator's prior work. And so it always defaults to nitpicking about details. I recommend that we rank individual categories [see Attachment D for handout that complements this presentation outlining investigator tracks]. We would take the five categories we currently have, modified slightly, and rank each independently. Beforehand, three primary reviewers would do the ranking and develop a score based upon that rank. If the reviewers agree it is their third grant out of the best, then it's clearly in the 30th percentile.

At the face-to-face meeting, we try to overcome the discrepancies between those scores, because some reviewer may actually have a good pile versus others, and that can have influence. So if we're going to do something like ranking individuals, particularly in the case of the investigator being one of the components, it does a disservice to the fact that some of our investigators have an outstanding track record. I suggest developing an award for outstanding investigators [see Track B in Attachment D]. In that track, the investigator does nothing more than submit his/her name as a candidate. If you're really outstanding, everyone knows who you are. Your name goes to three different study sections, for example. And those study sections will score based upon that name. You can call it a beauty contest if you want, but I see nothing wrong with that. I would suggest taking that pool essentially out of the study section and distributing it across NIH. This can help to mitigate the difficulties of variances in study sections.

Emerging Themes
Dr. Tabak

I'm going to share with you some emerging themes. Because I prepared this prior to this meeting, it will give you a bit of a check to see what your comments are like relative to the other comments we've been hearing. We hope it will reassure many of you that in fact we have been hearing what you have to say, because many of the comments here resonate with those of your colleagues around the country. However, no priority is implied in what I will share with you, and this is not a summary of what NIH plans to do. It is only meant to facilitate additional discussion.

There has been a lot of input about review criteria and focus, and the structure of applications. We've heard a lot today about the virtues, or lack thereof, of reviewing the project versus reviewing the person. Several of you have alluded to the fact that in the intramural program we do retrospective rather than prospective review.

Some have argued that reviewers should be blinded to applicants' identities during a first phase of review. And we've heard why two stages of review may or may not work. Yet, many will argue that issues related to environment or investigators per se, unless they're luminaries, don't really count anyway, so why go through that process.

Many folks have said that there's way too much emphasis on methodology, and in particular, too much emphasis on preliminary data. Some have argued this is killing innovative and risky ideas that typically will have minimal precedent.

There has been lots of discussion about reviewer mechanisms and mechanics. The wisdom of a crowd is something we have heard repeatedly. How do you reconcile the need for additional persons vs. the need to have a more focused, more intimate interaction? In part, that might be solvable through electronic or other virtual review mechanisms, and/or by two-stage processes.

We've heard a lot about the value, or not, of using the so-called editorial board model. Many have indicated the need to establish some sort of applicant-reviewer dialogue to correct factual errors during some initial stage of the review; whether we do that as part of a two-stage review or just integrate it into the present type of system doesn't seem to matter.

There has also been lots of discussion about interdisciplinarity. Can the editorial board model be used that way?

We have heard ideas about different types of review for different types of science. Should we do clinical research reviews together with basic science? Does clinical research require involvement of patients and/or their advocates in the review process?

We've heard issues related to community-based research. Should we involve members of the community in the review process? In earlier meetings, particularly the last meeting in Washington where we spoke to voluntary organizations, there was a great deal of resonance about that.

It turns out that clinical trials tend to be new submissions. I think you all are aware that A0s, the first submission, historically do very poorly at NIH; we fund under 10% of these. Are we inadvertently biasing ourselves against new clinical trials?

Interestingly, the only thing that has surprised me about this meeting is that we haven't heard a lot about SBIR/STTR. I thought that coming to this part of the world, where there's so much innovation and entrepreneurship, we would have heard about this. So the rhetorical question is, are academics the right folks to review small businesses? Some would argue that perhaps they're not.

The low A0 success rate clogs the queue. Some have argued that we should use a pre-application process to provide a rapid identification to separate truly noncompetitive ideas from those that might be competitive

Several of you alluded to the need to consider administratively funding those applications that contain minor and easily correctable deficiencies rather than sending them back to the queue. Many have spoken about the need to provide more useful feedback to applicants. Many would like to triage triage, in particular for new investigators. And several of you alluded to the need to unambiguously tell applicants if their applications are not recommended for revision and resubmission.

How do you maximize review and reviewer quality? How much context should reviewers on study panels have, and how much should be provided to them? Right now, we have a firewall at NIH: The review and program are separate. Some have argued that we need to empower reviewers by providing portfolio analyses.

How do we entice you to be reviewers? Should we add time to your extant applications? Should we supplement your extant grant to give you some part-time administrative help? Should we give you salary support? Several of you have alluded to the need to be more flexible with regard to service; would two times a year make more sense than the present three? Should we have mandatory service?

Should we provide in-depth training for reviewers? Lots of people have suggested this. How do we get folks to focus on strengths and not weaknesses? How do we get a focus on potential impact of the proposal and not methodology? Is the role of the study section to rewrite the proposal or to review the merit?

Should an ombudsperson be assigned to every study section? Some say we need to rate the reviewers. Some say we need to rate the scientific review officers. Some say we should anonymize the process completely, such as NSF does.

We've heard a lot of about ranking. One idea that has come up repeatedly is to revisit scores at the end. Some have said we have to completely redo the scoring system. Many of you have already alluded to the potential value of matrix scoring for different criteria or dimensions. What is most discouraging to applicants is failing to do better upon resubmission. One possible reason is a new set of reviewers with a whole new set of issues; another is that the first time around, nobody wanted to tell the applicant that the "baby is ugly."

Many have said we should reinstitute mechanisms that are unique to new investigators, either separately within the same panel, or by separate study sections. We have many mechanisms at NIH, and each institute tends to use them in a different way. Some have argued this is too confusing and that the numbers need to be reduced.

Other issues relate specifically to R01s:

- How many R01s are enough?

- Should we require a minimum percent effort from an R01 PI?

- Should R01s remain the gold standard of investigator success? The argument here is R01s vs. the top-down so-called big science.

Comments from Members of the ACD Working Group

Mary Beckerle, Ph.D.
University of Utah

I'd be very interested to hear comments on the following questions:

- How can we improve our ability to identify and support the most innovative science through our peer review process?

- How might we attract the highest quality reviewers to participate in the process?

- How can we maximize the effectiveness of the peer review process, given this enhanced interdisciplinarity of science?

- Do you have additional ideas about how we can most effectively support our young people and allow them to get fully engaged in the scientific process?

Bruce Alberts, Ph.D.
University of California at San Francisco; Chair, Boundaries Report

What kind of experiments would you suggest, and what are the controls for these experiments? For example, do you want the same grants reviewed by two different mechanisms to see how they compare? I am very encouraged that NIH is willing to experiment, because we don't want to change something dramatically and find we've made a huge mistake. We are, after all, scientists, so we'd like to get some data.

Compilation of Key Points Made during Open Discussion Sessions

Applicant Feedback/Interaction

- There has to be some mechanism involved whereby the applicant gets feedback on significance. If it is a grant that isn't going to make it based upon its merits alone, no matter if it were perfectly written and all the data were perfectly done, then that feedback must be provided. Maybe there has to be nothing more than a simple check box.

- The problem of lobbing back and forth between reviewers who may disagree is a significant issue. I've been told in study section that you need to review the current application, and if you don't think it's as good as the first one, then you need to score it that way. I think we should score on the best features of the application and give feedback to the PI that there was disagreement, and we leave it up to the PI to decide how they want to resolve it.

- In order to foster innovation, improve the quality of the review process, and overcome the hurdle of misunderstood applications, perhaps we could do something Robert Wood Johnson has done. They literally require you to be there for some of their grants and partake in a 10-minute interview. Applications cannot be adequately reviewed if the applicant does not have an opportunity to express his or her point of view to the reviewers.

Applications

- If grants focus on explaining what one has done in the last funding cycle and why it was important, and then project into the future without much detail what one wants to do, one can actually shorten applications and emphasize impact.

- Retain preliminary data. It lets me know what's been done. I also know the quality of the data and how well the investigator conducts research.

- The reliance on preliminary data is a serious problem that stifles innovation.

- Some have suggested setting aside the senior investigators as special cases. That is actually our present system, because the preliminary data become more and more important, and that just selects out the people who have the resources.

- If the NIH application process is shortened to 15 pages, which I generally think is a good idea, it will be a disadvantage to new investigators. One way around that is to allow a 5-page appendix for new investigators that would include the detailed methods they can't demonstrate through previous publications. For senior investigators, they cite the method because they developed it, and that's it.

- With 25 pages for R01s and 15 pages for R21s, the innovative ones are shorter. Presumably that means you should get some benefit of the doubt, but they don't give you that. I've had the comment, "I know you're limited to 15 pages, but I just wish I had more information to be sure you'd be right." So I think if you do cut down, you've got to understand what those unintended consequences will be.

- We thought the problems we have with NIH are because we are a small firm, but I see we are not alone. The preliminary study part of the proposal is killing us, because our experience shows we have to do almost half the work in advance. As a small business, that's very costly for us, and we don't know how to get around it.

Basic Research

- All the work we're doing now is based on extremely good, well-thought-out basic science. The investigators who did that work early on didn't know how it would be used. While I have a great deal of respect for some of the arguments made here today, we really have a need for good basic science research.

- I like the idea of elevating clinical translation through all phases of research, but I don't think that's true for all basic research. We don't know what's going to be the application of somebody's basic research.

Collaboration

- I chair the state of Florida's Biomedical Research Advisory Council, and we oversee about $20 million a year in peer review grants. We do experiments on the peer review process all the time, including, for example, numerical scoring of components of overall score. We would welcome the opportunity to collaborate with NIH. I'm sure that organizations like the American Heart Association, American Cancer Society, etc., that also run large-scale peer review grant programs would welcome the opportunity for peer review process improvement through collaboration with NIH. [Dr. Richard Bookman, University of Miami]

Criteria for Funding

- I recommend having more of a focus on broader impact, like the NSF model, and less on strictly biomedical relevance. NIH has a separate mechanism that typically funds smaller institutions with underrepresented minorities, called the Minority Biomedical Research Support (MBRS) Program. However, people on the MBRS panel have found misconceptions at NIH regarding the purpose of MBRS: that it is to fund teaching institutions and not to eliminate the disparity of underrepresented minorities who get into Ph.D. programs and eventually become investigators.

- What is the meaning of impact? Impact on the science the person is applying for, or on the mission and goals of NIH? How about significance? For what? It all depends on your viewpoint. Unless those terms are well defined, they mean different things to different reviewers. Once you clarify these, clinicians will become part of the effort; you'll find very few who are PIs anymore.

- Innovation by its very nature is going to be unpredictable. The risk for innovation has to be placed in the hands of the investigator and not in the study sections. Study sections, especially in times of tight dollars, are going to be inherently conservative, and they're going to act against innovation.

Electronic/Distant vs. Face-to-Face Reviews

- I favor the virtual review committees, or getting away from the face-to-face meetings. In writing, you can have an exchange of ideas and perhaps even consider getting the applicant involved. When you're in a face-to-face meeting, and you have to talk off the top of your head with only a few minutes to discuss an application, it really doesn't do the application justice.

- If face-to-face meetings could be replaced with conference calls, people would save time and money by not having to travel. I would also invite the applicant to stand by. They don't have to hear everything you're saying, but if you have a detailed question, you can invite the applicant to respond.

- Although I like the idea of online review and the fact that the more people we involve the better, we might want to consider how to track how ideas get borrowed, because more and more people will be able to see them.

- I propose “chat room reviews.” A proposal comes in, and an SRA assigns it to two or three reviewers, who submit their preliminary reviews with a score to a secure chat room to which the applicant also has access. Conversation goes back and forth in the chat room until a final decision is made. The ultimate decider is the SRA. This system allows correction of what I think is a very important problem with the peer review system: reviewer mistakes. This way, the expert on that application, the applicant, can help to correct these mistakes. The much-vaunted socialization of the study sections would be replaced by a much more productive socialization, completely focused on the science of the application itself. Then the scores would be manipulated by statistics to come up with sort of normalized scores and ultimately percentile ratings.

- I don't think the idea of “chat room reviews” would actually help a not-already-funded area of research.

- Face-to-face gives us the opportunity to have an open and very frank, focused discussion on a particular application.

- Although I don't like traveling several times a year to NIH as part of a review panel, I think it is important. And there is one critical difference between reviewing papers and grants. When you review a grant, you have to submit a score and stand in front of a qualified panel to defend it, which gives a level of credibility to the system.

- The study section should meet in person. However, we all know that the strong advocate, the strong spokesperson, the extroverted personality can dominate in a study section discussion. The order in which your grant gets reviewed also needs to be taken into consideration. The purpose of a face-to-face is to take the grants that have been scored prior to coming to study section that represent approximately twice the number of grants that can be funded, and for the study section to rank those grants 1 to 20, 1 to 30, or 1 to 40. Each person would get their scores on each of those four criteria from every reviewer, along with the justification for those scores, and then ultimately a ranking.

Expertise

- You can enhance the quality of the review of an application if you ensure it is reviewed by people with expertise or experience in your area of research. I do my research on in-prison substance abuse treatment programs, and it's clear from some of the summary statements that the reviewer has never been inside a prison, nor does that individual know very much about what it's like to conduct research within a prison. And it's very disheartening when you feel as though your score's been negatively impacted because the person, with all good intentions, really didn't know or understand what you were trying to do.

- My experience is that the review committees don't know the delicacy or complexity of multidisciplinary proposals, and that it’s not expected of them.

- Translational science calls for people with some experience and knowledge in both basic and clinical research studies to be able to capture that element.

Funding

- Regarding the “magical” 15% number, the process has gotten politicized so that large-scale piles of money are given to strong universities, like Washington University, that are mainly political in nature. I could name one university that gets $25 million that certainly does not produce 100 R01s worthy of production. This has to be looked at very seriously because it comes at the cost of individual investigator initiatives.

- There is no prorating for the size of the student population at different schools, so places like San Francisco State get a certain amount of funding that's a little too much to warrant certain kinds of funding mechanisms, but obviously not enough to maintain the level of quality of research that we want, and to train students to move into the system.

- I strongly discourage the idea we heard earlier about safer funding situations. One of the strengths I've seen in the American system is strong competitiveness. Whatever change is made, that needs to be maintained. How much funding you get should be limited only by the ability to do something with the money.

- Putting a limit on the number of grants an investigator can have, or the percentage of effort they have to invest in those grants, is a really good idea.

- NIH plans must include continuity of support as a major issue, because that's what keeps people's laboratories alive, and what forces them to have multiple grants so that in case one fails, they've got another rock to step on.

Incentives To Serve

- In terms of attracting the best reviewers, give us more time to write our grants after we're on the study section.

Indirect Costs

- I'm pretty sure that my overall amount of money being brought into the university is substantially more than what I'm getting in return. Is there some mechanism for the indirect costs to go directly to the investigator?

- Indirect costs are not profit margins for universities. There are very specific places they go, and I'd be happy to share with anyone in great detail exactly where the University of Washington's indirect costs go.

Junior Investigators

- I'm a new investigator, and I think I'm voicing the views of a lot of people in my department and in similar kinds of institutions. We feel a little left out of the NIH peer review process. R01s are particularly difficult to come by. Many of us believe the diversity of the panels isn't what it should be. We should focus on experts in the field and not necessarily how well-funded somebody has been.

- As a recently funded new investigator, I favor having separate study sections for new investigators. When you combine grants in the same study section from established labs composed of 20 or more members with someone just starting out who has only a few individuals in the lab, it's very hard to be unbiased.

- Give new investigators a new category. When my grant is in the same study section as my post-doc mentor, who is an academy member, and we're funding at a 9% rate, what are the chances I'll get funded? It's just not conceivable.

- There needs to be a separate study section for new investigator R01s. Even if R01s are discussed at a separate time during an existing study section, reviewers are still seeing applications with a lot more preliminary data, a lot more meat that these established investigators have had a long time to generate and pull together. The separate study section should be an R01, not an R29 type of mechanism, because one of the problems with the R29 is that it wasn't renewable. You want to encourage new investigators to have a continuous chance at getting funded.

- I'm not sure that creating a separate panel for new investigators is such a good idea. Ultimately, they have to learn to compete with the other applications. In our study section, we review all the new investigators all at once. I think that serves a purpose because we are cognizant of the fact that we're now reviewing new investigators and giving them the "new investigator discount."

- If an investigator is supporting graduate students rather than research technicians, shouldn't that earn us some brownie points in terms of renewal applications? Young scientists get engaged when you work side by side with them in the lab, and that isn't built into any NIH-funded funding mechanism.

- New investigators are frankly and clearly and absolutely underserved. Something has to be done specifically for new investigators, and it can't be a tweak to the system.

Junior Investigators: Grant Mechanisms

- The new investigator award mechanism is extremely important, and we need to bring it back.

- I want to speak about possibly going back to a mechanism specifically for junior investigators. I think we're losing so many brilliant young people well before they submit their first R01. I was one of those people funded on an R29. You had 4 to 5 years to get your lab up and going. It was a relatively small amount of money, but it was really what you needed at that point. We weren't talking big science, but getting a few people in your lab to do good work. They knew they were part of a group of people watching over them, where they could get mentorship. Now a lot of junior investigators are very discouraged at the graduate student level, and they are looking for other avenues to be creative with their scientific interests.

- Another program for young investigators that hasn't been mentioned and should be encouraged is the K99/R00. This is for post-docs, and it requires no preliminary results for the R01 funding. A key component is mentorship during the final 2 years of post-doctoral training. We don't fund enough of these.

Miscellaneous

- The solutions that are adopted have to decrease the amount of time put into proposals.

- The investigators who are really getting a squeeze are those at mid-level, and that is where we don't hear much about what is actually occurring.

- I don't think the scoring system is where you should deal with issues of underrepresented minorities. It's actually in the RFPs.

- The statement, “We want to fund the best science by the best people” – that's sometimes what NIH wants to do. At other times, they want to encourage particular fields or particular ways of doing research that may not be the best science by the best people, but may instead be different approaches or different settings. A one-size-fits-all doesn't work. I've seen that particularly in the SEPs

I've been involved in that are promoting, or looking at, community-based participatory research where there are specific mechanisms that demand specific things. And then you get to the review process that does not ask you to review any of those specific things that were called for in the RFA. They don't ask you, for instance, to look at community involvement or the quality of the partnership or things like that, except to the extent you can fit them in these four standard criteria and one single-score mechanism. So I'd encourage looking at having the shoe fit the particular goal of the particular program, and admit there really are different goals and different programs for different efforts by the NIH.

- Are there any lessons you can learn from the small business side? They have a handle on quite a few things – for example, when we write small business proposals, we are not required to have preliminary data. I think small businesses have industry representatives included in their study section. As a reviewer, you are required to submit your comments ahead of time. I think the SRA can pass some of the comments to the PI.

- I would like to argue against the proposal of having specific special investigators awarded on a separate track or mechanism based on prior results. People have risen to the point of being chairs of departments on the basis of research that is continuously NIH funded but that has never affected the life of a single patient. Instead, we ought to make sure grants are funded in accordance with their likelihood of affecting patient care.

Program Announcements

- If there is some room for translation, this can be taken care of in the program announcements. But currently, program announcements have no meat to them whatsoever. A program announcement with no dollars attached to it means the application just winds up getting scored by the study section with instructions not to score it in relation to what the program announcement is, but rather to score it in relation to the science. So maybe there need to be changes that will help the directed questions.

Public Participant Role

- Having a public participant role in the study sections should be an institutionalized feature. That's because when you demand and expect of the public representative the same quality of informed scientific and supportable opinion that everyone in this room subscribes to, you'll get it.

- Research that affects communities should have community input in the review process. I support encouraging more community-based organizations and community members to be allowed in the review process, and encouraging training to allow community members to more fully participate.

- I don't think consumer advocates would work well for all study sections, especially for some of the basic research issues.

Reviewer Comments

- I believe most of us have received reviewer comments that we feel are unfair and uninformed. We might consider having the reviewers identify themselves and base their comments on actually citing literature why they disagree with a certain point. If you make a scientific judgment, you have to be able to stand by it.

- There should be some mechanism for going back after the discussion to try to make the reviews reflect what the discussion addressed, as opposed to the separate reviews at the very beginning.

Scoring/Ranking

- The suggestion of the speaker to rank independently on creativity and innovation, etc., is desirable, and those rankings may very well be different. The problem I see is that each study section serves many institutes, and a low ranking might be the highest ranking for a given institute. So I think the rankings would have to be combined at the study section meeting face-to-face, and then separated by institute. This system would work well just so long as the study section is for an institute alone, but not across institutes.

- Ranking should be done at the end of the study section.

- I would like to suggest a scoring system that could be combined with a ranking system in which we actually separate out quality of the application, impact of the significance of its expected outcomes, the experimental design and likelihood of success, and the novelty. Is this person leading or following the pack? This could be methodology, new paradigms, or new translational evidence. If each one of these were individually scored 1 to 10 by all reviewers, you would get a total. The person whose grant is being reviewed would see all those scores combined and would get a sense right away whether to resubmit that grant. By breaking it down and having reviewers justify their numerical score for each one of those categories, that information can be transmitted directly to the investigator, and they could see exactly where their grant stood.

- I want to put in for the ranking independent of the group process. I find that the senior persons on study sections actually do lead what happens, and there are very good social experiments that show how that works.

- When you rank something, you have to hold everything in short-term memory in order to compare the different ones. And when we get much past 10 units or proposals, we run into trouble. But there is a technique that's been developed by

William Stephenson called the Q-sort technique. I strongly urge that you look at that, because it will allow you a system for ranking a large number of items.

- I like the idea of ranking, but I think it would be a disaster for new investigators unless they were placed in separate study sections.

Staffing Panels

- Study section membership should be made more flexible without decreasing quality – perhaps decreasing the length of time that reviewers serve, and/or decreasing the number of times per year they go to study section.

- It would be very beneficial to have a member who sits on a particular study section perhaps go to a different study section once a year so his/her expertise gets distributed around.

- At the moment, it is sort of implicitly expected that if you've received substantial NIH support, you should pay your dues and serve on study section. I think that should be an explicit expectation, so that if you have received substantial support and are an associate-level professor or up, you are expected to respond to an invitation to review.

- Community members should be included in the upper layers of the review process. So, for example, I've heard about a two-stage process whereby the first stage is scoring, and the second stage is overall consideration of significance. I think this piece in particular should be something a community representative should be a part of.

- I'd like to advocate for more racially and ethnically diverse membership in order to include members from the communities that are often targeted by research.

Study Section and Review Models

- There is a need for crosscutting evaluations for research from multiple disciplines. Often an ad hoc reviewer is brought in, but it's usually totally inadequate, and that person gets dominated by everyone else.

- I know that the investigator is important, and the institution is important, but I suggest that the review be conducted in a blinded fashion. I've sat in too many study sections where the name, weight, and cronyism of the person who submitted the grant is clearly rewarded.

- I propose two guidelines for use when reviewing applications. First, does the proposal addresses an important question, and is the specific aim worth funding? If the answer is no, there is no reason to go any further. The second guideline would be whether the design could address the main aim. Oftentimes, reviewers lose the forest for the trees, forget about the importance of the question, and come up with other, irrelevant questions.

- It may be better to organize peer review in a way like the MRC in England, where the grants are sent out and external reviewers write the critiques. Then the study section members are put in a very different psychological position. Here, they nitpick little details, whereas in the other system, they are the wise judges of the critique that is delivered by the outside reviewers. Also, the person who is reviewed has a chance to rebut.

- The current triage system is a very bad idea. It doesn't give appropriate feedback to the PIs, institutions, department chairs, and so forth, who are responsible for mentoring the PIs. It doesn't save much time on the part of the reviewers, as they still need to read the applications and think carefully about them.

- I have experienced what seems almost a pilot of what we have discussed, and that was to review the NIH Director's new investigator board. I got 35 grants, 10 pages each, and 2 weeks to review them and write a very succinct report, using 5-point scales in three categories. Out of that came a fairly natural ranking system. From my point of view, it worked really well.

- The "trust me" approach, where senior investigators submit their own name, has a lot of merit to a limited extent. The problem is, it may well foster excellent biological discoveries, but not necessarily translational and clinical science that makes a difference in human health. That could easily be solved by adding modules, or incentives, for people to go to the clinic or go into patient populations beyond just the traditional mechanistic thinking.

- I'm concerned that in the reorganization of the study sections, the focus on organ systems and disease processes excludes those who study agents (e.g., drugs, environmental chemicals). If we maintain the current system, there needs to be some consideration for looking broadly at the existing structure of IRGs, with an eye toward adding new IRGs or reorganizing.

- There is a general perception among surgical subspecialists that the reorganization process CRS underwent 5 years ago to more disease-specific lines has disadvantaged investigators from those subspecialties, in that the grants are now unbundled and distributed to different IRGs and to different study section within the IRGs. Grants that are translational and clinical then end up with inadequate numbers of actual peers to give the grants adequate reviews. Compounding that is the problem of getting adequate numbers of peer reviewers from the clinical subspecialties because of their workloads. My proposal is to bundle surgical subspecialty applications into targeted study sections.

Training/Mentoring

- Think back to the first time you were on study section. I was one of these up-and-coming new investigators with his first grant who thought he knew what was going on. So I clanged these grants left, right, and center, nitpicked the whole bunch of them, until I learned 2 years later that this was not the thing to do. It would have been better had I been trained not to do that in the first place.

- If NIH held conferences to allow the post-docs, even the Ph.D. students, to get training during their early stage, they would figure out if they have the ability to become independent researchers and good scientists. Many scientists don't know how to write grants.

- I was involved in a survival skills courses for post-docs that was especially directed toward minorities and minority institutions. The course, funded by NIH, helped people learn how to write grants and papers, and give presentations. These skills are very important.

- I attended a NIDA technical assistance workshop entitled “Substance Abuse, Criminal Justice, and HIV in African Americans.” The purpose was to try to promote successful applications for minority applicants in this particular area of research. NIDA invited a number of experienced researchers as well as newer investigators such as myself to share our experiences. We were assigned as mentors to two or three minority applicants to review their concept papers and provide feedback. This workshop can serve as a model for other NIH institutes to develop some sort of national mentoring network to help people who want to become new investigators.

Closing Remarks

Dr. Alberts

I know a former graduate student who succeeded with the young innovator award after trying for an R01 several times. When I asked whether her last unfunded R01 should have been funded, she said, "No." When I asked why she submitted it, she said she was told by senior people to only submit things you know are safe and that you know you can do. And so at least by that one example, the young innovator award was very successful in allowing her to actually say what she wanted to do, and having reviewers look at it from a different perspective perhaps than they normally would. I know that NIH has a lot of data, and I'd like to see that analyzed in great detail to see how we might build on that experience. One of the things I encourage NIH to follow is the vast majority of people who should have been funded and weren’t, but they put together interesting ideas. What happens to those people later? Do they go into R01s? Are they going to get funded for more innovative things than they otherwise would have tried?

I’m very concerned about young investigators, because we're going to lose the next generation. I am in contact with young people at UCSF, and they're thinking of all kinds of other careers, which is not necessarily bad, but we want people to go into research, too.

Dr. Beckerle

Bruce's comment about this young colleague is something that must be happening around the country. If we can collectively come up with experiments and ideas that support and nurture the creativity of our young people, that's going to be extremely important.

There was a lot of discussion today about preliminary results. One important point is that preliminary studies have a value to the extent that they show that a hypothesis is well grounded. But when it's taken to the point that it shows you know the results of the proposed work, it's gone too far.

Dr. Yamamoto

I am chair of the Board on Life Sciences for the National Academies, which gave rise to the committee (chaired by Tom Cech) that generated the K99/R00 proposal. I urge you to look at the committee’s report *Bridges to Independence: Fostering the Independence of New Investigators in Biomedical Research*. It includes a lot of proposals focused on support for new investigators that we would love to see at least an enhanced conversation about within the community, perhaps leading to experiments and changed policy.

I'd like to look at a couple of ideas we have heard as a way to point out the complexities we are up against and to say that it’s true, one size does not fit all, and one solution is not going to work either. We will need a number of changes from a number of different vantage points.

For example, different suggestions were made about providing guidance to the investigator on how to improve the grant application. But it was also pointed out that the next time around, two new reviewers might review it and find more things they consider even more damning, and the score goes down after the investigator conscientiously responded to the first critique. And so one of the complex facets of the coin is the question of whether the peer review system should be a mentoring exercise, or should it simply be to make a judgment about the scientific merit of the application? Mentoring is critical, especially for young people, but maybe it should be carried out by colleagues of whomever is writing the application, whether junior or senior.

There is a lot of merit to the idea of separate study sections for new investigators. NIH has been back and forth on this with things like training grants. Should new investigators be in their own separate study sections, or should they be embedded within the R01 study sections? On the down side, you might say separation is a terrible idea, because what we all want is to ensure that the right people are in the room who know the area, have the expertise, and can appreciate the details and nuances of what is presented. If you bundle all the new investigators together, the chances of having that kind of coverage go down, not up.

John Featherstone suggested that maybe you don't want to make sure you have focused expertise, but instead people who understand the issues for new investigators and who can be generalists about science. That raises this whole other issue of whether we are disadvantaging new investigators by shortening the grant applications, for example. What's so sacred about 25 pages? Every young person we train and hire has gotten their job by writing 2- or 3-page applications. Maybe we should depend on the experience that says you can make a call on the prospect of a young person and feel pretty confident that you're making the right one. This is what we do when we hire assistant professors, after all. So that is the other side of that coin as well.

Additional points include the following:

- We heard ideas for increasing efficiency in a number of areas, including transmitting the right information from the applicant to the reviewers, getting pre-meeting feedback on preliminary critiques that are online, and establishing an online review system that draws from a national pool of registered reviewers that can draw direct feedback from the investigator.

- Regarding face-to-face meetings, we heard different perspectives that were passionately defended. I think this is something we will have to pay a lot of attention to, and perhaps will be the substance for some very good experiments.

- A couple of very interesting ideas were put forth regarding outstanding investigators, including a special sort of grant that would provide $300,000 for perhaps extended periods of time

- Other ideas include giving new investigators a small amount of money to supplement their startup costs and so forth, and paying attention to that middle ground of the people who are struggling to get conventional R01s.

- Some have raised the question of whether NIH supports transformational work that crushes paradigms and makes us restart our thinking. Are there ways to reach for that? Do some of these ideas for supporting outstanding investigators do that, or is it just another way to support really productive people?

We owe you a great debt of gratitude for coming, participating, thinking, and speaking in the way that you have. Thank you.

Attachment A: Handout Prepared by Presenter Dr. Peter Bacchetti—Annotated References

Attachment B.1: Handout Prepared by Presenter Dr. David Busath—Traditional Research Career Trajectory

Attachment B.2: Handout Prepared by Presenter Dr. David Busath—Proposed Standard Research Trajectory
Attachment C: Handout Prepared by Presenter Dr. Ken Dill—Rating NIH Proposals: Ranking vs. Scoring
Attachment D: Handout Prepared by Presenter Dr. Fred Schaufele—Investigator Tracks

Attachment A: Handout Prepared by Presenter Dr. Peter Bacchetti

Annotated References

Impact of sample size on the projected scientific and/or clinical value of a study

1. Bacchetti P, Wolf LE, Segal MR, McCulloch CE. Ethics and sample size. *Am J Epidemiol*, **161**:105-110, 2005.

This introduces the idea of diminishing marginal returns as sample size increases and discusses how this invalidates previous ethical condemnation of so-called "underpowered" studies.

2. Bacchetti P, McCulloch CE, Segal MR. Simple, defensible sample sizes based on cost efficiency. *Biometrics*, in press.

This establishes diminishing marginal returns for a wide variety of measures of projected study value that have been proposed in connection with sample size planning. It also proposes and justifies methods for choosing sample size based on cost considerations, without a need to try to estimate power. It shows that a method tailored for innovative studies outperforms the conventional approach while avoiding the vexing difficulties that innovators currently face when trying to justify a sample size.

Difficulty of calculating sample size

3. Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives Of General Psychiatry*, **63**: 484-489, 2006.

This shows that even a favorable situation for sample size planning, having relevant preliminary data from a pilot study, leads to unreliable sample size calculations.

4. Vickers, AJ. Underpowering in randomized trials reporting a sample size calculation. *Journal of Clinical Epidemiology*, **56**: 717–720, 2003.

This systematically examines a seemingly best-case situation, pivotal randomized clinical trials published in 4 leading medical journals. It nevertheless finds that most used sample size assumptions that were off by enough to produce >2-fold errors in the sample size calculations. About a quarter were off by >5-fold.

Manipulation of sample size calculations to produce desired results

5. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results, *Annals of Internal Medicine*, **121**:200-206, 1994.

This described such manipulation as a "sample size game".

Sample size and peer review

6. Bacchetti P. Peer review of statistics in medical research: the other problem. *Br Med J*, **324**:1271-1273, 2002.

Fallacy that p>0.05 proves no effect

7. Rumbold AR, Crowther CA, Haslam RR, Dekker GA, Robinson JS. Vitamins C and E and the risks of preeclampsia and perinatal complications. *NEJM*, **354**:1796-1806, 2006.

This found a reduced rate of infant death or other serious outcomes, with the reduction corresponding to needing to treat 39 women for each outcome prevented. It nevertheless concluded that "Supplementation with vitamins C and E during pregnancy does not reduce … the risk of death or other serious outcomes".

Bias against innovative studies

8. http://cms.csr.nih.gov/NewsandReports/ReorganizationActivitiesChannel/BackgroundUpdatesandTimeline/FinalPhase1Report.htm

This report from an NIH panel on peer review noted that "an obsession with preliminary data discriminates against bold new ideas, against young scientists, and against risk taking. For new ideas, little or no preliminary data may be required." This fine statement, however, does not help innovators with the practical problem of trying to somehow prove that their studies will have 80% power.

Attachment B.1

Handout Prepared by Dr. David Busath

Traditional Research Career Trajectory

Research Career Support Scale (Stages => Levels)

Stage 1: Setting up a lab: years 1-3
Stage 2: Making a story; making a reputation: years 4-6
Stage 3: Attracting students; expanding the lab: years 7-10
Stage 4: Creating a multifaceted research program: years 11-16

Stage 5: Blossoming of research program: years 16-20
Stage 6: Production: years 21-35
Stage 7: Semi-retirement: years 36-70

	Pubs Per year	Research Focus	New Instruments (~$20,000)	PI Faculty Position	UGS	GS	Tech	Post Doc	People-Meetings per year	Annual Cost Salary Supplies Travel
Yrs 1-3	1	Narrow	1	Asst. Prof.	1	0	0	0	1	$21,500
Yrs 4-6	2	Open 2nd area	1	Asst.Prof.	2	1	0.5	0	3	$63,500
Yrs 7-10	3	Expand 2nd area	1	Assoc.Prof.	4	2	1	0	5	$109,500
Yrs 11-15	4	Add 3rd area	1	Prof.	6	2	1	1	7	$175,500
Yrs 16-20	4	Expand 3rd area	1	Prof.	7	2	1.5	1	10	$211,000
Yrs 21-35	5	Develop 4th area	1	Prof.	8	3	1.5	2	12	$275,000
Yrs 36-70	1	Several areas	0	Emeritus	2	1	0	0	2	$22,000

Traditional Research Career Trajectory

Standard	Pub-Proposals per year	$K per Proposal	Pubs per year	Citations per year	$K Per Pub	$K Per Citation	Annual Program Income	Level Up Instrument ~$20,000
I, years 1-3	1	20	1	2	5	5	$35,000	1
II, years 4-6	1	22	2	4	6	6	$58,000	1
III, years 7-10	2	24	3	6	7	7	$111,000	1
IV, years 11-15	3	26	4	8	8	8	$174,000	1
V, years 16-20	3	28	4	8	9	9	$192,000	1
VI, years 21-35	4	30	5	10	10	10	$270,000	1
VII, years 36-70	1	20	1	1	8	4	$32,000	0

Proposed Standard Reward Trajectory

Attachment B.2
Handout Prepared by Dr. David Busath
Proposed Standard Reward Trajectory

A New Contract with Researchers

- ✓ NIH-supported positions associated with faculty positions
- ✓ Long-term commitment to support continuity
- ✓ Rewards for Productivity and Development
- ✓ Complementary Rewards for Competitive Goals (Pre-calculated)
- ✓ Rewards for High-Impact work (Post-calculated)

Strategic Approach

- *Tiered reward plan*
- Online submission of 1-page "Pub-Proposals"
- *Base funding for pub-proposals* is paid upon approval, after online judging with ~90% approval
- *Reward funding* for papers and citations is based on quality and paid over 2-3 years, providing a steady stream of program income
- Study section determines *level advancement* based on a) proposal fulfillment rate; b) publication rate; c) proposal submission rate; d) citation rate; e) proposal, publication, and citation quality; f) instrument request

Tactical Approach

- Proposals judged by 3 disinterested-author/interested-keyword (DAIK) judges, with each score normalized by the judge's score-history
- Publication quality assessed by journal reviewers. NIH-supported journals are those that allow reviewer scores and names to be shared confidentially with NIH for initial publication quality evaluation. On-line registered volunteer reviews are also accepted and factored into compensation.
- Publications credit the most germane proposal for proposal-fulfillment tracking
- Citation quality is determined automatically, with 10% reduction for collaborator (joint author with any of the authors in the past 3 or 5 years); 10% enhanced for abstract, introduction, results, or discussion citations.
- Proposal judges and publication judges work on-line for a nominal wage.
    - Enhance general readership & cross-fertilization
    - Provide side-job opportunities for published scientists

Attachment C: Handout Prepared by Presenter Dr. Ken Dill
Rating NIH Proposals: Ranking vs. Scoring

Our study sections currently rate grants by *scoring* each one from 100 to 500. If, instead, each reviewer were to *rank-order* his or her grants, say from N (best) to 1 (worst) – not a huge departure from how we often review grants in other settings – I believe it could solve some significant problems for NIH.

Problems this could solve: (1) We would not need for NIH reviewers to have face-to-face meetings in Bethesda, which is costly and time-consuming. (2) SRAs have a hard time spreading out scores, to choose among near equals; scores tend to cluster around the mean, for simple math reasons. (3) The current face-to-face review system is conservative, tends toward fault-finding, and wastes valuable reviewer time in listening to conversations unrelated to his/her principle expertise.

The basic problem: These problems arise largely from a simple math problem: we need all the reviewers in the same study section to be normalized to a single reference point. This is the reason our reviewers currently must spend two days in a room hearing about, and making up their own scores, for proposals they haven't read. However, this normalization problem can be solved by replacing our scoring system with a ranking system. For the purpose of illustration, I assume below that we first shorten R01 grant proposals, say to 10 pages. I think this is important to do anyway, but here, it also keeps the math simple.

Example of a simple ranking system: Each reviewer reads 20 proposals (at 10 pages each, this is still less burden than the current 10 proposals x 25 pages for each reviewer). Each reviewer performs a triage step: he/she gives the worst 10 proposals a zero. The reviewer then scores the first good proposal (worst ranked proposal) with a 1; the next better proposal with a 2; … the second-from-the-top proposal with a 9, and the very best proposal with a 12 (in this example, giving two extra points because it's the best one). In short, each reviewer rank-orders his/her proposals, rather than just scoring them. Now, provided that each proposal is read by multiple reviewers, each reviewer can work in private, and a pocket calculator can do the normalization straightforwardly.

Suppose we have 8 grants to review. Suppose the "perfect rankings" are given by the right-hand column:

Grant Number	Ranking
B	1
D	2
F	3
H	4
G	5
A	6
C	7
E	8

We have 4 reviewers: a, b, c, and d. Each Reviewer gives 3 zeroes (unranked) and scores 1, 2 and 4 from worst to best: 6 total reviews. The matrix below shows a possible scoring from the 4 reviewers.

Grants	Reviewers				Total
	a	b	c	d	
A	0	-	0	0	0
B	4	-	4	4	12
C	0	0	-	0	0
D	2	4	-	1	7
E	0	1	0	-	1
F	1	2	2	-	5
G	-	0	1	0	4
H	-	0	0	2	2

Even with some disagreement among reviewers, this scheme correctly ranks proposal B highest, D second best, F third best, and H fourth best, as required, even though only one reviewer is "perfect". It's fairly immune to noise.

To show that it is robust, suppose reviewer d tried to "blackball" grant #2, by giving grant #2 a 0 and grant #3 a 4, instead of the scores above. In that case, the rankings would still correctly put #2 highest, #4 second best, and #6 third best. Only the lower-ranked grants would be rearranged. Hence blackballing doesn't affect the best grants, in this ranking scheme.

The reviewing burden in this system is not onerous. Suppose you have P proposals; suppose you want each proposal to have m reviews; and suppose you want each reviewer to perform r reviews. Then the number, R, of reviewers you need will be R = mP/r .

In a typical case, if you have P = 100 grants on a cycle, and you want m = 5 reviews per grant and if each reviewer reads r = 20 proposals, then you need R = 25 reviewers on that cycle, about the same as in the current study section system.

Such a ranking system has some additional desirable properties: (a) that any one reviewer can effectively advocate for a grant (if extra points are given for the top scorer or two), giving it a particular boost (favoring innovation, I believe), but (b) cannot blackball a grant. Also, because proposals are ranked rather than scored, this system normalizes out some reviewer biases, such as the difference between cranky and Pollyanna reviewers. In addition, such a ranking system is fairly flexible. Suppose a reviewer cannot decide which of two proposals is better, say the best and second best. Rather than scoring 12 and 9, those two proposals would now both be scored with the average value, 10.5. This gives a small boost to two good grants, rather than a big boost to one grant, sending a sensible signal within the scoring.

ATTACHMENT D
Investigator Tracks

	Track A (Most Investigators)	**Track B** (Self-Identified Outstanding Investigators)
Submissions -problem: outstanding investigator	Complete Proposal	Name Only
Funding -problem: comparing applications of different scope	1, 90% of funds 2, No More than Full Modular Budget	1, 10% of funds 2, Automatic 1.5x Full Modular Budget
Review Scoring -problem: study section dynamics (reviewer differences) (temporal differences) (niche scoring) (new investigators)	Full Review 1, Three Reviewers 2, Initial Score based on Rank in each of five categories: A, Significance B, Study Design C, Innovation D, Capability (investigator and environment) E, Likelihood of success 3, New Investigator receives at least a 3rd place equivalent rank in Capability 4, Study Section discusses non-triaged applications 5, All study section members provide rank score in each of the five categories 6, Aggregate score tabulated automatically from category scores	1, Scored by All Study Section Members Prior to Meeting 2, Awardees submit specific aims afterwards 3, Study Section reviews budget only

Attachment D: Prepared by Dr. Fred Schaufele, University of California, San Francisco