In this Issue

- Implementing New Rigor and Transparency Policies in Review—Lessons Learned
- Why Do You Review? Tell Prospective Reviewers via Our Blog
- Statisticians Share Insights for Applicants and Reviewers
- Pilot Results: Scoring with an Expanded Half-Point Scale
- Online Briefings for Applicants and Reviewers
- New Policies on Application Appendices and Post Submission Materials
- Predictive Nature of Criterion Scores on Impact Score and Funding Outcomes

Implementing New Rigor and Transparency Policies in Review—Lessons Learned

After scientists at NIH in various fields and pharmaceutical companies raised serious concerns about the reproducibility of NIH research, NIH responded by launching multiple efforts to enhance the rigor and transparency of the research it funds. A big one involved changes in review criteria that we implemented in the last round of grant application reviews that involved 525 meetings and nearly 10,000 reviewers.

So How Did It Go?

**Overall, reviewers rose to the challenge.** In premeeting trainings, SROs emphasized that changes were needed in the wording and substance of reviews. NIH asked reviewers to consider four new emphases:

- Scientific premise
- Scientific rigor
- Consideration of relevant biological variables, including sex
- Authentication of key biological and/or chemical resources

Most reviewers made explicit notes on these factors in their critiques, and the topics came up in panel discussions.
But work remains. Not everyone got the message. NIH continues efforts to make it clear that it has elevated the degree of attention that must be paid to sex as a biological variable, and also to resource authentication. Even those reviewers who have always thought carefully about scientific premise and rigor should reflect on the design and methodological considerations that are critical for work in their field. Some reviewers thought about the new emphases but used old language in their critiques, which made it harder for SROs and program officers when dealing with large numbers of applications.

"Premise" caused confusion. NIH intended for reviewers to consider the scientific foundation of the proposed work. That is, reviewers should critically ask whether the studies or preliminary data leading to the proposed work are scientifically sound. Although this sounds like an obvious consideration, multiple studies, covering multiple fields of science show that scientists have often overestimated how replicable published work is—even when published in top journals. Performing this review of premise can be intellectually demanding and many reviewers did a great job. Others confused “premise” with “scientific significance” or discussed whether the hypotheses of the study were reasonable. Significance refers to the importance of the study; premise refers to its scientific foundation. While a weak premise clearly undercuts the potential significance of a proposal, a strong premise (empirical foundation) does not necessarily make a study significant.

"Sex" was hard to talk about, sometimes. NIH expects that sex as a biological variable will be factored into research designs, analyses, and reporting in all vertebrate animal and human studies unless there is a compelling scientific argument for not doing so.

The impact of this policy varies considerably across different areas of science. Panels were challenged at times to really sort through—

- What is convention, and what is good science?
- What is adequate incorporation of sex in study design?
- Can inclusion of both sexes actually reduce scientific rigor, for example by increasing physiologic variability?
- If a disease affects one sex predominantly but not exclusively, is that sufficient justification for single sex studies?
- What should investigators do with sex specific data that is not sufficient to investigate sex differences beyond reporting it?

There are no blanket answers to such questions at this point, so reviewers need to bring their best scientific thinking to the table in order to achieve the fundamental goal of achieving replicable, generalizable science.

However, answers to many frequently asked questions about rigor and transparency can be found on the NIH grants website.
Reviewers know a good authentication plan when they see it. Reviewers know that “obtained from a trusted source” is not an authentication plan. Some applicants still need to be educated on that point. And some fields lack consensus guidelines for crucial resources. However, reviewers—both experienced and expert in their fields—had clear ideas about whether or not the one-page plans were adequate.

Why Do You Review? Tell Prospective Reviewers via Our Blog

When recruiting reviewers, our Scientific Review Officers try to explain why others in the scientific community are willing to serve on study sections. But those of you who have served could better explain why you do it. We encourage you to use our comment feature below to answer this question. Perhaps you will inspire those who haven’t reviewed to volunteer or say yes when we ask.

Statisticians Share Insights for Applicants and Reviewers

Some members and staff of the American Statistical Association (ASA) recently held discussions with senior CSR staff. We covered many important topics and insights. We thought you would be interested in the key points, including how statisticians can (1) help improve rigor and reproducibility, (2) be key members of a research team, (3) identify common statistical issues, and (4) play valuable roles in peer review meetings.

Rigor and Reproducibility

“One of the biggest reasons for research that is not reproducible is the degree of confidence in the result is hugely overstated,” said Dr. Marie Davidian, William Neal Reynolds Professor, Department of Statistics, North Carolina State University. “This happens . . . because the design may have been bad, there may have been too much data snooping in the analysis planned, or the methods may not have appropriately captured all the sources of uncertainty.”

How Statisticians Can Be Key Members of a Research Team

ASA members encourage investigators to engage a statistician at the outset of a research project and fully integrate them into their team. “We can often
operationalize the scientific method, so an investigator will produce a scientific hypothesis and the statistician will turn it into a testable hypothesis,” said Dr. Karen Messer, professor of biostatistics at UC San Diego. “We can help applicants quantify what they know, but more importantly, quantify what they don’t know and put the quantified measure of uncertainty on what comes out of the study. That ultimately is what makes science reproducible.” Dr. Davidian summed it up, saying “Statisticians bring clarity, focus, and specificity to a project.”

Common Framing and Design Problems in Applications

Our discussion with ASA members continued with detailed examples of the problems statisticians can help investigators to avoid or bring to light in reviews:

- **A Poorly Defined Hypothesis:** One of the great things a statistician can do is help analyze “the inferential chain” that is at the heart of a research application, said Dr. Messer. “In other words, is the question well stated? Are the measures appropriate? Is the population of interest well-defined and appropriate? Is there control of the false-positive rate?”

- **Incomplete Description of Data Collection:** “In many grant applications, it is impossible to tell what data points are actually being collected and when/how they are measured,” said Dr. Andrew Althouse, summarizing the thoughts of his ASA colleagues. He is a biostatistician at the Magee-Women’s Research Institute at the University of Pittsburgh Medical Center. “It should be abundantly clear exactly how many patients/samples are being collected, if they are being measured just once or at multiple time points.”

- **Dealing With Missing Data:** “Frequently there are issues of missing data [such as subject dropouts] . . . how they would impact the analysis . . . and how they will be handled,” said Dr. Davidian.

- **Power Analysis and Sample Size Fails:** A group of ASA members listed four common problems:
  - Power analysis is not aligned with the data analysis plan. For example, when the PI powers for main effects when effect modification is of interest, or powers for an ordinary hypothesis test when an equivalence test is most appropriate.
  - Power calculations are inaccurate. For example, they might be based on inflated estimates from a small pilot study.
  - Power calculations are unclear and vaguely described.
  - No power calculations are provided at all.

Common Problems with Data Analysis Plans

- **Too Simplistic:** “Some analysis plans are very vague and ignore current state of the art methodology,” said Dr. Althouse. “They instead propose outdated or inappropriate analysis methods.”
• **Too Complicated:** “Conversely,” he continued, “some applicants propose the most sophisticated analysis possible when a simple comparison of conditions is called for.”

• **Not Suitable:** “The analysis plan may not be suitable for the primary aim,” said Dr. Althouse.

• **Ignoring Important Things:** Sensitivity analyses and multiple comparisons.

### Advice to Statisticians Serving on Review Groups

**Get valuable experience:** ASA members encourage statisticians first to get experience working on collaborative teams, as it will help them to better appreciate the perspectives of applied researchers and reviewers. Even if they don’t have this experience, they should really consider serving on a review group that is at least broadly aligned to their area of interest. Immersion is really the best way to learn how peer review works.

**Meet great colleagues and learn a lot:** Serving as a reviewer is a valuable way of finding new collaborators and datasets. “You can actually learn a lot,” said Elizabeth Stuart, professor of mental health, biostatistics, and health policy management at Johns Hopkins Bloomberg School of Public Health. “You can develop great friendships with colleagues . . . by serving on panels, which is useful in a lot of different ways.”

**Draw on your strengths:** Statistician reviewers need to “trust themselves and their training,” said Dr. Stuart, “and identify and articulate problems they see.” According to Dr. Messer, they should “. . . be a little skeptical on the areas where we’re not such experts, and be willing to find or accept reasonable compromises in our areas of expertise.”

**Enjoy:** “I find it a lot of work,” said Dr. Messer, “but it’s very stimulating and enjoyable. I think you learn a lot!”

### Check Out these Great External Resources

- [Ten Simple Rules for Effective Statistical Practice in PLOS Computational Biology](#)
- [FASEB’s Statistics in the Life Sciences: A Summer Reading List](#)
- [American Statistical Association Website](#)
Pilot Results: Scoring with an Expanded Half-Point Scale

The current NIH application scoring system gives reviewers nine score choices (1-9). In practice, many reviewers only use a portion of this scoring range in their initial scoring and during their meeting discussions, where the majority of applications are scored in the 2-4 range. In other words, reviewers frequently only use three of their scoring choices to evaluate the best applications, and the best score (1) is being sparingly used by most panels.

This can contribute to score compression in which a high percentage of applications are scored in a fairly narrow range. For example, a moderately compressed study section could score more than 25% of its applications with a score between 30 and 10. This problem may be exacerbated by low funding levels. In study sections with severe compression, competitive applications are often scored within only two score choices (1 and 2).

In addition, a review of scoring patterns across NIH has revealed peaks in the distribution of final overall impact scores at 20 and 30, indicating many tied scores. Score compression and ties make it difficult for program officers to distinguish between the very best applications reviewed in study sections, particularly when several applications receive identical scores and/or percentile ranks within the same study section.

CSR designed a pilot study to test whether allowing reviewers to vote in 0.5 point increments, based upon the score range set by the assigned reviewers, can reduce score compression and ties by giving reviewers more flexibility in score choices following discussion of applications.

Pilot Design

The pilot was run alongside the normal review process at study section meetings in the 2016/05 and 2016/10 council rounds. Thirty-three study sections in the 2016/05 and 11 study sections in the 2016/10 council rounds participated.

Following the discussion of applications, the assigned reviewers set the range with integer scores following standard guidance. All reviewers entered integer scores for the official scoring of applications based on the score range unless indicating that they intended to vote outside the range.

Reviewers were provided with separate score sheets that included a column for unofficial half point scores ranging from 0.5 points below their official score to 0.5 points above the official score. For example, if the score range for an application was 2-3, the allowable range for final scoring became 1.5 – 3.5. If a reviewer had scored an application with a 2 officially, the reviewer could enter 1.5, 2.0 or 2.5 in the half point column; for a score of 3, the reviewer could enter 2.5, 3, or 3.5.
A score of 1 remained the best possible score; a score of 0.5 was not allowed. Half point scores 0.5 lower or 0.5 higher than the score range did not require reviewers to identify themselves as voting outside the range. If reviewers had already identified themselves as officially voting out of the range, they could enter a score either 0.5 lower or 0.5 higher than the official score they entered.

**Results**

Aggregated analysis of the data revealed that, when reviewers used the half point option, they were more likely to raise the score of an application (for example, a move from 2 to 2.5) than to reduce it (a move from 2 to 1.5).

Scores calculated from half point data were compared to the official scores for 1,371 discussed applications from 39 study sections. The percentage of scores at 20, 30 and 40 was reduced when reviewers were given the half point option, as shown in the figure below. Score compression was also improved for a few study sections that were highly compressed based on the official scores.

![Distribution of Application Scores: Original vs Half Point](image)

**Reviewer Survey Found Strong Support for a 0.5 Scale**

Surveys were sent via email to 311 reviewers in the 11 study sections participating in the 2016/10 council round to gain feedback on the half point pilot. Reviewers were asked to complete the survey near the end of the meeting. They were assured that their response was voluntary, identities would not be disclosed, and only aggregated responses would be used in analysis.
Completed surveys were received from 138 reviewers, 118 of which used the half-point increments during final scoring. Overall, a majority of reviewers indicated that the option of providing half-point increments improved their ability to prioritize applications based on impact and that the resulting scores more accurately reflected the scientific merit. Two thirds of surveyed reviewers agreed or strongly agreed that NIH policy should be changed to permit half-point increments in scoring.

Tell us what you think by submitting a comment below!

Online Briefings for Applicants and Reviewers

CSR will host three online briefings for applicants and reviewers in November and December 2016:

- Fellowship grant applicants
- R01 grant applicants
- Basic research grant applicants and reviewers

<table>
<thead>
<tr>
<th>Briefing Title</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>8 Ways to Successfully Navigate NIH Peer Review and Get a Fellowship Grant:</strong></td>
<td>November 2, 2016</td>
</tr>
<tr>
<td>This briefing will cover the key things applicants need to know about the submission and review of their fellowship applications.</td>
<td></td>
</tr>
<tr>
<td><strong>8 Ways to Successfully Navigate NIH Peer Review and Get an R01 Grant:</strong></td>
<td>December 1, 2016</td>
</tr>
<tr>
<td>This briefing will cover the key things applicants need to know about the submission and review of their R01 applications.</td>
<td></td>
</tr>
<tr>
<td><strong>NIH Peer Review Briefing for Basic Research Applicants and Reviewers:</strong></td>
<td>December 2, 2016</td>
</tr>
<tr>
<td>This briefing will reaffirm NIH’s commitment to basic research and help applicants and reviewers do their part in proposing and reviewing basic research.</td>
<td></td>
</tr>
</tbody>
</table>

All of the briefings will run from 2:00 to about 3:00 p.m. Eastern Time, including a 30 minute Q&A period.
How to Participate in the Briefing

- **Go to** [www.csr.nih.gov/webinar](http://www.csr.nih.gov/webinar) **to register for the briefing you wish to join.** You will not need to download special software. You will just need a reliable Internet connection and browser.
  - Register by October 31 for November Briefing
  - Register by November 30 for December Briefings

- **Submit questions for the Q&A session before or during the briefing** by sending them to the moderator at AskExperts@csr.nih.gov.

- **Go to** [www.csr.nih.gov/webinar](http://www.csr.nih.gov/webinar) **on the day/time your briefing is scheduled.** Click on the link that will be provided there.

View Archived Briefings and Past Webinars

- **View archived webinars and PowerPoint slides on our webinar web page:** Meet the Experts in NIH Peer Review webinars for R01, R15, Small Business and Fellowship grant applicants as well as a webinar for university research administrators.

- **View our 2016 Briefings and PowerPoints about a month after broadcast on our webinar web page.**

New Policies on Application Appendices and Post Submission Materials:

The use of NIH application appendices has been decreasing for some time, but they still create problems, which led NIH to further limit the types of materials that can be put in them starting January 25, 2017.

Many applications containing appendices have had to be withdrawn because the appendices contained noncompliant materials (such as unpublished manuscripts, figures, tables and other data, or experimental methods). If appendices with such materials somehow made it to review, they could give the applicants an unfair advantage over other applicants who followed the rules.

The new appendix policy was designed to make the rules more simple and clear, to make the review process more fair for everyone, and to decrease the number of applications withdrawn for noncompliance.
Significant Change

Accepted manuscripts and non-publicly available papers (or publications of any kind) will no longer be allowed in an appendix, but news of an article accepted for publication since submission of the application will be allowed as post-submission materials.

Other Important Things to Know

- Applications may still include appendices with informed consent/assent forms and blank surveys, forms, and other data collection instruments (as appropriate). In addition, applications containing clinical trials may include the clinical trial protocol and Investigator’s Brochure from an IND application, as appropriate -- unless alternate instructions are provided in the funding opportunity announcement (FOA).

- Unless a FOA requires certain information to be included in the appendix, reviewers are not required to consider the material in their review.

Learn More by Reading the new appendix policy in the NIH Guide.

Change to the Post-Submission Materials Policy

A new NIH Guide notice on post-submission materials mainly consolidates and clarifies current policy.

Significant Change

“News” of a new publication is now specified as the List of Authors/Institutional Affiliations, Title of article, and Journal or citation (if available).

Predictive Nature of Criterion Scores on Impact Score and Funding Outcomes

What are the key criterion scores that drive impact score and funding outcomes? Staff from the NIH Office of Extramural Research recently published new research that further explores this question. After examining 123,000 competing R01 applications, they published their results in PLOS-one. In short, they found that the Approach and, to a lesser extent, the Significance criterion scores were the main predictors of an R01 application’s Overall Impact score and the likelihood the application will be funded. To learn more and participate in the
ongoing discussion, visit the “Open Mike” blog on the NIH Office of Extramural Research Web site.

For Another Perspective: See a paper recently published by NIH staff: NIH Peer Review Scored Review Criteria and Overall Impact. The lead author, Dr. Mark Lindner, has since been named Director of CSR’s Office of Planning, Analysis and Evaluation

Subscribe to Peer Review Notes: www.csr.nih.gov/prnotes
Send comments or questions: PRN@csr.nih.gov

Center for Scientific Review
National Institutes of Health
U.S. Department of Health and Human Services