Simplifying Review Framework: Feedback from the Request for Information

National Institutes of Health

April 28, 2023

Executive Summary

NIH is proposing a revised, simplified framework for peer review that will reorganize the major regulatory criteria (42 C.F.R. Part 52h.8) under three factors and reduce the number of non-score driving review considerations that peer reviewers evaluate in research project grant (RPG) applications during the first stage of review. NIH believes that these changes will allow peer reviewers to refocus on the critical task of assessing scientific merit and will improve those assessments by reducing bias.

From December 8th, 2022 – March 10th, 2023, NIH solicited feedback on the simplifying review framework proposal. This report summarizes the feedback received; the views expressed in this report are of the respondents of the request for information (RFI) and are not necessarily an endorsement by NIH of these statements or suggested actions. All feedback received was reviewed and considered; this report presents the major themes and/or feedback contained in the comments. There were 780 unique responses received from individuals, 30 received from scientific societies, and 30 received from academic institutions. Responses were considered whether they were submitted through the RFI webform, in response to related Review Matters and Open Mike posts, or via email. The report contains three sections: overarching themes, individual responses, and scientific society/institutional responses.

Overarching Themes

The majority of all individual and scientific society/institutional respondents to the RFI favored the proposed changes. Generally, it was felt that the simplified changes were likely to achieve the goals of refocusing reviewers on the elements that are truly important in identifying the most impactful research, through reduced burden, better emphasis on key evaluations, and reduced impact of global reputation of the investigator and environment in favor of a focus on evaluation in the context of the research proposed.

Respondents also discussed current issues in interpretation and application of peer review concepts and criteria and requested greater attention to these in the new factor-specific definitions. For example, respondents called for expanded definitions of *Innovation*, clarity around *Feasibility and Rigor*, and more consideration around the definition and assessment of the investigative team as a whole. Respondents noted that well-crafted definitions will further aid reviewers in making unbiased reviews.

Many respondents, whether individuals or societies/institutions, commented on the need for NIH to develop strong, effective training for reviewers and clear communication to the external community regarding the proposed framework and scoring system. It was felt that the proposed changes will be successful only if reviewers understand the new criteria and are well-equipped to apply it.

A smaller subset of respondents presented concerns about the proposed changes:

- Some disagreed with the proposal to make *Factor 3: Expertise and Resources* an unscored factor (but still contributing to overall impact score).
    - One subset of these respondents argued that the *Investigator* and *Environment* are critical to the success of the proposed work, and therefore, a factor score should be assigned.
    - A separate subset of respondents appreciated the intention of this change, but argued that removing criterion scores would make it more difficult to detect bias that may be influencing the overall impact score.
    - A slightly smaller subset of respondents argued that the proposed changes did not go far enough to address potential bias in peer review and advocated for a blinded or partially blind review process.

- There were some comments that the framework is merely cosmetic and will be ineffective in meeting the stated goals of the plan.
- There were some concerns that reducing the number of criterion scores would increase subjectivity and ambiguity in informing the overall impact score or lead to score compression.

Methods

From December 8th, 2022 – March 10th, 2023, NIH solicited feedback on the simplifying review framework proposal. There were 780 unique responses received from individuals, 30 received from scientific societies, and 30 received from academic institutions. Responses were considered whether they were submitted through the RFI webform, in response to related Review Matters and Open Mike posts, or via email.

Comments from multiple sources were combined into one dataset for analyses. Responses submitted by scientific societies or academic institutions were then identified and separated out into a second dataset, however a similar coding approach was used for both datasets. Analysts first read each comment in its entirety and coded its relative support at a high level (Favorable, Unfavorable, Mixed Sentiment, Neutral (SRF-related), or Not Applicable). Comments that were tagged "Not Applicable" discuss other topics outside of review criteria and the proposed plan were not included in this analysis. Analysts then used a coding scheme to tag subtopics in each comment based on components of the new framework (e.g., comments relating to a specific criteria or factor, or Additional Review Criteria), features of the plan (e.g., changes to Factor 3 scoring), or the principal drivers of change (e.g., bias in review, reviewer burden/reviewer responsibilities). Analysts worked in close collaboration during the coding process to revise the coding scheme iteratively to accommodate new themes that emerged, and to achieve adequate inter-rater reliability.

Feedback from Individuals

3-Factor Framework and Scoring

Most respondents were supportive of efforts aiming to simplify the review of RPGs and shared that this framework allows reviewers to focus on the scientific and technical merit of the grant. Those in support of the new framework shared the following viewpoints:

- The organization of criteria and consolidation of scoring for *Factor 1* (*Importance of the Research*) and *Factor 2* (*Feasibility and Rigor*) is viewed to reflect and codify the features of a proposal that are most score-driving in current practice and reduce the conceptual overlap and redundancy between the current criteria (e.g., *Significance* and *Innovation*).
    - Similarly, the coupling of *Investigator* and *Environment* criteria for *Factor 3* (*Expertise and Resources*) introduced synergy between the two criteria, and the categorical assessments proposed were viewed as guiding the reviewer to appropriately consider *Investigator* and *Environment* in the context of the proposed work.
- Creating *Factor 3* (*Expertise and Resources*) as a non-scored criterion was viewed by many respondents as likely to help reduce the effects of reputational bias in peer review by structurally shifting the focus onto the scientific and technical merits of the proposal. These respondents felt that attempting to score *Investigator* and *Environment* on the current 1-9 scale is an arbitrary endeavor and leaves room for bias.
    - Respondents were optimistic that these changes would create a more equitable playing field for investigators and institutions during peer review and hoped to see this reflected in review and funding outcomes.
- The 3-factor framework was perceived as likely to reduce reviewer burden and to facilitate clear and effective discussion and critique-writing.
- The removal of some administrative criteria was overwhelmingly welcomed as a strategy to reduce reviewer burden.

A smaller group of respondents were against the proposed changes to the peer review framework or held the viewpoint that the proposed changes would do more harm than good. Key criticisms of the proposed plan include:

- There are concerns that the 3-factor structure is merely a cosmetic reshuffling of existing review criteria which is viewed to be ineffective in addressing reviewer burden and would not help improve peer review quality. These respondents perceived the proposed changes as a large adjustment with little payoff.
- There are concerns that reducing the number of criterion scores from 5 to 2 would increase ambiguity and subjectivity in informing an overall impact score.
    - Relatedly, there are concerns that the reduction of criterion scores will contribute to score compression by making it more difficult to distinguish and rank applications.
- There are two main concerns with the proposal to remove criterion scores for *Factor 3* (*Expertise and Resources*):
    - Some respondents felt that the categorical assessments offered were too limited to comprehensively assess whether the investigative team's expertise and credentials were able to support the proposed project. Respondents felt that considerations of track record and past productivity were meaningful in assessing feasibility and likelihood of successfully executing the proposed project and should affect the overall impact.
        - More respondents felt more strongly that *Investigator* should be scored over *Environment*, but some respondents maintained that institutional environment also speaks to feasibility.
    - A separate group of respondents felt that by removing criterion scores for *Investigator* and *Environment*, it would be more difficult to detect and mitigate reputational bias that may still be influencing the overall impact score. Furthermore, some respondents maintained that reputational bias or other types of bias will continue to impact assessments under *Factor 1* and *Factor 2*. Examples of how this might occur include in critiques of the perceived significance of the topic or research area, subjective assessments of feasibility to accomplish the proposed research, or nitpicking the approach of less-well known investigative teams. These respondents expressed a desire for attention to other types of bias in review, examples provided focused on career stage, underrepresented racial/ethnic groups, and gender.
        - Some respondents shared that assessments that are not scoreable but are able to contribute to an overall impact score (such as *Factor 3* or *Additional Review Criteria*) cause ambiguity to which factors most contributed to informing the overall impact scores. Some respondents thought that these unscored sections should not be able to impact the overall impact score.

Some respondents shared mixed sentiments about the proposed plan—while most acknowledged the problems of reviewer burden and bias in peer review and expressed a desire to see these issues addressed at NIH, some felt that the outlined plans would either be ineffective in meeting stated goals or do not go far enough. There are also respondents who supported some aspects of the plan, but not other aspects. The following mixed sentiments occurred most often:

- Some respondents shared that they appreciated a move to a simplified review structure but felt that *Factor 3* needed to be scored.
- Regardless of the level of support for the proposal to change *Factor 3 (Expertise and Resources)* to an unscored criterion, many respondents expressed that adopting a blinded review model would be the only or most effective way to combat bias in peer review and level the playing field for investigators and institutions. While the blinded models described had slight variations, often considerations of *Investigator* and *Environment* would only be considered after the first two factors are evaluated.
    - A smaller set of respondents discussing blinded review models were against blinded review, stating that considerations of *Investigator* and *Environment* were critical in assessing the project's likelihood in completing stated goals, or spoke to the impracticality in blinding the applications themselves (self-references of preliminary research, risk of reviewers recognizing the investigative team based on other information described in the application).
    - Alternatively, a few respondents would like to see considerations of *Investigator* and *Environment* to be removed from the first stage of review altogether.

Factor 1: Importance of the Research

Most respondents were in favor of combining *Significance* and *Innovation* criteria under *Factor 1: Importance of the Research*. Those in support of this factor conveyed that the distinction between the two criteria is often arbitrary. Respondents felt that it needed to be made clear to reviewers that *Significance* and *Innovation* do not need to be weighed equally when assessing *Factor 1* and felt that combining both criteria into one factor affords this flexibility.

Most respondents were in favor of combining *Significance* and *Innovation* criteria under *Factor 1: Importance of the Research*. Those in support of this factor conveyed that the distinction between the two criteria is often arbitrary.

A smaller group of respondents had reservations about combining *Significance* and *Innovation* criteria into *Factor 1: Importance of the Research*. These respondents felt that *Significance* and *Innovation* are independent measures and combining them into a single factor would make it difficult to inform *Factor 1* scoring for proposals that are found to be significant but not innovative (more often the case), or innovative but not significant. There are concerns that this will reduce transparency and increase subjectivity in informing Factor 1 scoring.

A few other considerations were shared by a handful of respondents, including:

- A few respondents also felt that *Innovation* would be better coupled with *Approach (Factor 2)*.
- There were a few respondents who were against the use of "Importance" in the titling of *Factor 1* and asked for alternative titling to be explored.

Significance

Of the two criteria, *Significance* was more often considered to be more important in evaluating the overall merit of the grant.

However, some respondents argued that *Significance* is subjective and dependent on the positionality of the reviewer and thus, vulnerable to biases on whether a research area or topic was considered significant or important by the reviewer. These biases may disproportionally impact research performed or valued by underrepresented or historically-excluded groups. These respondents felt consideration should be given as to how biases could be mitigated within the definition of *Significance*.

Innovation

This criterion was viewed to be the more problematic of the two, regardless of whether the respondent supported or disagreed with the plan to combine *Significance* and *Innovation* under *Factor 1*. Respondents felt that technological or methodological innovation was currently favored by reviewers over other types of innovation, such as conceptual, theoretical, or proposals where existing methods are applied in a new way. This narrow interpretation of *Innovation* was also viewed to hurt the prospects of a proposal from under-resourced labs or institutions, which may not have access to the newest or most advanced technologies.

Innovation was viewed to be a problematic criterion. Respondents felt that technological or methodological innovation was currently favored by reviewers over other types of innovation. Respondents hoped to see a wider definition of *Innovation* considered under factor definitions, supplemented with guidance and training for reviewers.

There was a view among many respondents that valuable and significant research that may use appropriate and well-established methods are currently underappreciated by reviewers, and there is mixed sentiment as to whether combining the *Innovation* criterion with *Significance* will help or exacerbate the problem.

- Those who are in favor of considering *Significance* and *Innovation* together think it will afford reviewers flexibility in weighing the two and are hopeful that review and funding outcomes for applications proposing important research utilizing standard methods will improve.
- Others are concerned that *Factor 1* scores for these types of applications will be dragged down by a narrow definition of *Innovation*, negatively impacting their review and funding outcomes. A handful of respondents felt that *Innovation* could be dropped entirely, with the view that a proposal may not necessarily have to be innovative to be significant or impactful.

Respondents hoped to see a wider definition of *Innovation* considered under factor definitions, supplemented with guidance and training for reviewers.

Factor 2: Feasibility and Rigor

Many respondents are in support of *Approach* being a stand-alone Factor and acknowledge that this section is often a primary driver in informing overall impact scores. Those in support of the new framing of *Factor 2* as *Feasibility* and *Rigor* felt that it may lead to more focused assessments, rather than nitpicking the approach or a preference by the reviewer for other methods than the ones proposed. Much like *Factor 1*, respondents felt it was crucial to make clear to reviewers that there is flexibility in the relative weighting of *Rigor* and *Feasibility* in informing *Factor 2* scoring.

However, some respondents expressed confusion on how to interpret and evaluate *Rigor* and *Feasibility* and requested additional training and guidance in this area. Some wondered how assessments of feasibility will interface with other factor definitions, such as the capability of the investigative team, which would be assessed within Factor 3.

While factor-specific feedback was limited for *Factor 2*, a handful of other concerns were shared:

- A few were concerned that *Feasibility* will be interpreted as a requirement to provide extensive preliminary results that would need to be produced before submitting with a proposal, inadvertently resulting in conservative research being awarded.
- A few respondents felt that *Feasibility* and *Rigor* should be assessed separately.
- A few felt that *Feasibility* would be better coupled with *Factor 3* assessments.
- A few respondents felt that the focus on and definitions for *Rigor* and *Feasibility* were too limiting.
- A few respondents worry how this framework will further amplify an overemphasis on *Approach*.
- Due to existing confusion around *Rigor* and *Feasibility,* some respondents requested alternative titling for this factor be explored.
- It is not clear how Sex as a Biological Variable (SABV) will be considered in the new definition.

Factor 3: Expertise and Resources

Feedback related to changes in the scoring of *Factor 3* has been primarily shared in the first section. Other feedback related to *Factor 3* definitions is shared here:

Investigator

- There were mixed sentiments shared among respondents about assessments of "productivity" in current peer review practices.
    - Some respondents felt that the proposed definitions of *Investigator* under *Factor 3* do not provide room for assessments of productivity or track record and felt that this assessment was important. These respondents shared concerns that without adequate consideration of past productivity review outcomes would favor those who are skilled at grant-writing.
    - Others felt that productivity is a source of bias and is a subjective assessment, for example which journals the work is published in. A few respondents pointed out that assessments of productivity is also viewed to impact some groups more than others in the post-COVID research landscape.

- With the move towards multi-disciplinary research and multi-PI grants, some respondents thought *Investigator* definitions should reflect the need to assess the investigative team as a whole, including the capability to assess overlap in expertise within the team.
    - Related, there were some calls for greater attention to the diversity of the investigative team in assessments. A few respondents noted that considerations of the diversity and positionality of the team extend to other assessments as well, such as *Significance* and *Innovation*.
- There were also concerns with the language used in the binary assessments that may further exacerbate bias or is harmful, for example the subjective assessment of an investigator as "not fully capable". It was requested that alternate language be explored.

Environment

Environment-specific feedback was limited; however, the following feedback was shared:

- Assessments of *Environment* should reflect the specific facilities, resources and equipment needed to complete research aims, and away from institutional-level features that are extraneous to the proposed project. Respondents hoped that application instructions for resources and facilities can reflect this focus.
- Current definitions of *Expertise and Resources* focus too much on capital needs, rather than informational needs.

Additional Review Criteria and Considerations

Most respondents overwhelmingly welcomed the removal of certain review considerations (*Applications from Foreign Organizations, Select Agent Research,* and *Resource Sharing Plans*) from peer review and felt the change would decrease reviewer burden. However, some respondents felt that it was important for resource/data sharing plans to be evaluated by first-level review – as it relates to reproducibility and the importance of sharing data and resources across the scientific community. These respondents felt that reviewers are well equipped to evaluate proposed plans for resource sharing.

There were various other proposals and opinions put forward on which other assessments in the additional review criteria/considerations sections should be included in first-level peer review, and which should be able to influence or not influence the overall impact score. A common theme in the feedback requesting to remove certain sections from first-level review is that:

- These sections detract from the reviewer's ability to focus on the scientific merit of the grant; furthermore, it was perceived that these sections are not usually given sufficient attention in reviews.
- These assessments may be duplicative efforts on the reviewer's part when there are other entities responsible and better suited for assessing these components (i.e., IRB for Human Subjects Protections).
- Some assessments should be assessed just in time by NIH staff for applications scored in the high-impact range or applications slated for funding.

For the additional review criteria required to be assessed by peer reviewers, respondents requested more attention to reviewer training to guide appropriate assessments, such as those of *Human Subjects' Protections, Biohazards,* and *Budget*.

Implementation Considerations

In addition to the factor-specific training areas described above, respondents requested general training, guidance, and more information available for the following areas:

- Clear guidance, rubrics, definitions, and examples of how to appropriately assess and interpret each factor. Relatedly, respondents shared that active involvement by scientific review officers (SROs) will be needed during the review process to ensure that new factor definitions are adhered to, including intervention when biased assessments are presented. A handful of respondents shared some suggestions:

    - One suggestion was to produce a video that would be available for reviewers to reference, explaining each factor in detail.
    - Another suggestion was that pre-meeting reviewer training by the SRO should incorporate examples related to the area of science to explain how each factor should be evaluated.
- Respondents called for more guidance and information on the scoring system, both for reviewers and better communication to the extramural community. This includes the following information:
    - How overall impact scores and percentiles are calculated.
    - Which factors/additional review criteria and considerations will impact the overall impact score, how impact scores and percentiles are calculated.
- Guidance on how to appropriately consider *Investigator* and *Environment* in the context of the overall impact score, and strategies to avoid bias during review assessments.
    - Additionally, there was some confusion about how to assess Factor 3 within drop downs – some respondents were unsure how to select between "fully capable" and "appropriate", or "appropriate" from "concerns". That confusion likely resulted from how the information was presented in the RFI.
- Guidance for reviewers on identifying and distinguishing "score-driving" factors from "minor points", and how those should be considered in formulating criterion scores/overall impact scores. In addition, respondents requested more transparency in how reviewers arrived at their scores during discussion (framing reviews to identify score-driving factors) and greater attention to score-text matching.

Other implementation considerations include:

- Many respondents suggested reordering the application image to mirror the new framework: present the scientific information in the front of the application, such as the abstract/specific aims page, research strategy, references, followed by information such as biosketches, facilities and resources, budget, etc. Those in favor of this suggestion believe that it will strengthen the reviewer's ability to focus their review on the scientific and technical merit of the proposal – the current ordering may predispose reviewers to reputational bias on the investigative team or environment before they read through and consider the proposal itself.
    - A few respondents requested applicant instructions to more clearly reference how each section would be reviewed, and to revise the instructions for the research strategy to mirror the new framework.
- There were a handful of requests to pilot the new review framework before widespread adoption.
- There was a suggestion that instead of using "Major Score-Driving Factors" as a header or separate section from "Strengths" and "Weaknesses" – instead, only have three headers: "Score-driving strengths", "Score-driving weaknesses", and "Minor points".
- There was a suggestion to move *Factor 3* to "*Additional Review Criteria*", as it will now follow similar scoring rules as the other assessments in that section (will not receive individual criterion scores but will be able to affect the overall impact score).

Finally, there were requests for monitoring, evaluation, and eventual reporting of review outcomes related to implementation of the new framework. A handful of suggestions include:

- Defining and conveying clear metrics for success and a process for measurement of defined goals.
- Studies on factor scoring and review and funding outcomes.
- Track preliminary overall impact scores against measures of Environment and Investigator prestige at the individual and study section level.
- Case studies of comments in written critiques, such as how innovation is interpreted, and the degree to which reviewers' comments map on correctly to the framework and new factor definitions.

Clinical Trials and Human Subjects Applications

While the present RFI focuses on the review framework for non-clinical trial RPGs, a few comments were submitted about the review of clinical trial and human subjects applications. Respondents felt that Human Subjects and Clinical Trials Information (HSCTI) forms created a significant amount of reviewer and applicant burden and produces

redundancy in the information presented in the Research Strategy. Furthermore, respondents conveyed that the information presented in these forms do not improve the quality of review or influence the assessment of the overall scientific and technical merit. These respondents felt that these forms could be removed from review or provided just in time to NIH staff for proposals that are slated for funding.

Feedback from Scientific Societies and Academic Institutions

3-Factor Framework and Scoring

Most societies/institutions were supportive of efforts aiming to simplify the review of RPGs and shared the view that this framework emphasizes the scientific and technical merit of the proposal. Those in support of the new framework shared the following viewpoints:

- The revision and streamlining of the framework ensure that the process is more focused on the review criteria that drive the overall impact score, particularly around the approach and importance of the research. This is perceived to provide a thorough review which is focused on the scientific and technical merit of the application.
- The revised framework will promote equity and minimize potential institutional and reputational bias.
- The new framework reduces the burden and inconvenience for reviewers by reducing the number of elements for reviewers to consider and simplifies responses with drop-down ratings.

Many societies/institutions noted concerns regarding the proposed changes. Key criticisms of the proposed plan include:

- A handful of societies/institutions noted concern regarding the potential for score compression with a reduction to two scored categories. Respondents voiced concern that this could lead to clustering of applications with similar scores, without clear differentiation between proposals. However, respondents acknowledged that this may be mitigated through the anticipated reduction in administrative burden on reviewers with the new system, allowing reviewers more time to evaluate applications holistically.
- A few societies/institutions noted that it is unclear how the proposed simplification will mitigate undue influence of reputational and institutional bias as they are still considered within the overall application score. It is perceived that categorizing *Factor 3* as unscored but having it still influence the overall impact score makes the score less transparent.
- One society/institution noted that it was unclear what type of analyses was conducted on the current set of criteria, what the findings were, and how it ultimately informed the proposal that NIH put forward.

Some societies/institutions shared mixed sentiments about the proposed plan – while most recognized that issues of reviewer burden and bias exist in peer review and expressed a want to see they issues mitigated by NIH, some perceived the proposed plans would be either ineffective in addressing these issues or not going far enough. The following mixed sentiments occurred most often:

- Some societies/institutions noted that as currently written, simply reducing the criteria and removing criteria scores are unlikely to accomplish the intended goals of the effort, and there may be downstream consequences as a result of such changes.
    - Many societies/institutions expressed that adopting a blinded review model would be the only or most effective way to combat bias in peer review and level the playing field for all investigators and institutions. While the blinded models described had minor variations, often considerations of investigators and environment would only be considered after the first two factors have been evaluated without information on investigator and environment.
    - Several societies/institutions noted that while many of these efforts will help mitigate bias, they do not think it will stop the undue influence of reputational and institutional bias and believe that NIH could make more substantive changes in this area.
    - A handful of societies/institutions noted that it is not apparent how the proposed changes will translate to a noticeable reduction in review burden.

Factor 1: Importance of the Research

Most societies/institutions were in favor of combining *Significance* and *Innovation* criteria under *Factor 1: Importance of the Research*. Those in support of this factor conveyed that the proposed factors provide room for proposals that do not apply any novel approaches, but which nevertheless address critical knowledge gaps in a field. Out of the two criteria, *Significance* was more often considered to be more important in evaluating the overall merit of the grant.

Most societies/institutions were in favor of combining *Significance* and *Innovation* criteria under *Factor 1.*Those in support of this factor conveyed that the proposed factors provide room for proposals that do not apply any novel approaches, but which nevertheless address critical knowledge gaps in a field.

A small subset of societies/institutions had mixed viewpoints on the coupling of *Significance* and *Innovation*. A handful of concerns were shared:

- Some societies/institutions raised concern that it is unclear which of the factors takes precedence in the combined score – it was noted that research can be extremely significant and important but utilize already established methods, and therefore not be innovative.
- A handful of societies/institutions raised concern that *Importance of the Research* was perceived as a vague title with not much definition tied around how to evaluate the criteria.
- A handful of societies/institutions raised concern that reputational bias could persist in the scoring of *Factor 1* or be disproportionality applied to the overall impact score.

Several societies and academic institutions offered recommendations to NIH in relation to implementation considerations for *Factor 1*. The recommendations shared include:

- Many societies/institutions requested that NIH provide guidance and training for reviewers to highlight components of the application that contributed towards deriving the overall impact score.
- A handful of societies/institutions requested clear guidance on the relative weights of weighing *Significance* and *Innovation* in *Factor 1*. It is recommended that NIH provide clear, comprehensive guidance for reviewers on how to score this factor especially given that many reviewers are familiar with scoring *Significance* and *Innovation* separately under the current framework.

Factor 2: Feasibility and Rigor

A large majority of societies/institutions were in support of approach being a stand-alone factor and acknowledged that this section considers the prospect of whether the research project proposed as a whole entity is likely to be able to be accomplished and yield useful data, not whether individual components alone are likely to succeed. Some societies/institutions also noted that a focused section on *Feasibility* and *Rigor* will encourage reviewers to focus on the quality of the research methods and make the section easier to review.

A small subset of respondents had concerns associated with *Factor 2*. A handful of concerns were shared:

- A few societies/institutions noted that reviewers might couple *Feasibility* with *Factor 3* assessments.
- A few societies/institutions noted that reputational bias could persist in the scoring of *Factor 2* or be disproportionality applied to the overall impact score.
- A few societies/institutions thought that coupling *Feasibility and Rigor* was perceived as problematic as the two do not always align well and having one score for two items was viewed as confusing.
- One society/institution did not support the renaming of *Factor 2* and believed that it would have the unintended consequence of directing reviewer attention to assessing granular experimental details rather than focusing reviewer attention on the overall scientific approach to properly address a scientifically meritorious question and objective.

The following recommendation was shared by a society/academic institution to NIH as a consideration for implementation for *Factor 2*:

- Provide guidance and training for reviewers to highlight components of the application that contributed towards deriving the overall impact score.

Factor 3: Expertise and Resources

A majority of societies/institutions applauded the coupling of *Investigator* and *Environment*, the removal of a numeric individual score but ability to influence the overall impact score. It was perceived that this change would alleviate both institutional and reputational bias towards well-established investigators and highly funded institutions as well as reduce disparities in review outcomes. Those in support of the new framework believed that the coupling and non-scored nature of *Investigator* and *Environment* enables reviewers to focus on the scientific merit of the application rather than on preconceived notions or reputations of the investigator and institution.

A majority of respondents applauded the coupling of *Investigator* and *Environment*, the removal of a numeric individual score but able to influence the overall impact score. They believed that this change enables reviewers to focus on the scientific merit of the application rather than on preconceived notions or reputations of the investigator and institution.

A small number of societies/institutions had concerns associated with *Factor 3* and believed that the concerns would render the factor less valuable when it came to transparency in scoring. The concerns shared include:

- Some societies/institutions thought that a non-numerical score for *Factor 3* would render current bias invisible as it would not be evident in scoring.
- Some societies/institutions were confused as to how reviewers would use the proposed drop-down menus to evaluate Investigator and Environment. The confusion likely resulted from how the information was conveyed in the RFI. The RFI did not clearly indicate that the proposal is for two drop down menus – one for Investigator and one for Environment.
- A minority of societies/institutions believed that *Factor 3* should be scored and that the track record of the investigator ensures that the proposed study has the intended impact.

Several societies and academic institutions offered recommendations to NIH in relation to implementation considerations for *Factor 3*. The recommendations shared include:

- Some societies/institutions noted that as *Factor 3* is not individually scored but will be assessed and considered for the overall impact score, there may be challenges in this assessment, which could influence objectivity. Having clear guidelines on how *Factor 3* should be assessed and considered, as part of the overall impact score will be important to minimize bias.
- Several societies/institutions noted that NIH should include a supplemental section where investigators are required to outline how their proposed team, approach, trainees, and participants will advance broadening participation and inclusive excellence in biomedical research.

Additional Review Criteria and Considerations

Most societies/institutions that commented on additional review criteria and considerations welcomed the removal of certain review considerations (*Select Agent Research, Applications from Foreign Organizations*, and *Resource Sharing Plans*) from peer review. Many felt that this would aid refocusing attention to the scientific merit of the application and decrease reviewer burden.

Several societies/institutions offered recommendations to NIH in relation to implementation considerations for *Additional Review Criteria and Considerations*. The recommendations shared include:

- One society/institution recommended that NIH needs to further clarify for reviewers that *Additional Review Criteria* will be considered for the overall impact score, and to clarify for reviewers within the final framework why expertise and resources have been designated as *Main Review Criteria* rather than *Additional Review Criteria*, as this factor will not be scored individually yet will influence the overall impact score.
- There were a few recommendations about the inclusion of the *Inclusion of Women, Minorities, and Individuals Across the Lifespan*. One society recommended more training and attention by reviewers to this section as to not de-emphasize the importance of this section. Another society noted that more attention and training should be provided on how to score this assessment – "fully adequate" should only be utilized if the proposal has a specific plan to ensure that the demographics of the research population matches the demographics of the disease or condition being studied.
- One society/institution noted that NIH should consider eliminating the need for detailed budgets at the proposal stage, requiring them only for those projects selected for funding to focus the review on the scientific merit of the proposal as opposed to the potential price.

Implementation Considerations

In addition to the factor-specific considerations described above, societies/institutions requested general training, guidance, and more information available for the following areas:

- Many societies/institutions strongly encouraged NIH to develop robust training for reviewers, study section chairs, SROs, and applicants prior to launching the new framework to socialize the new criteria. Resources such as interactive trainings, rubrics, and other guidelines will be crucial for consistency in review and reduced administrative burden.
    - Some societies/institutions requested guidance on how to differentiate major score-driving factors from strengths and weaknesses, and what constitutes minor points. Additionally, it was noted that overall clear guidance and transparency is needed on how to weigh the factors in order to reach the overall impact score.
- Several societies/institutions recommended to develop overall guidance for scoring for how to translate the criteria into the 1-9 scale in an effort to be more transparent about how scores are derived, including out of range scores, and to prevent score compression.
- A few societies/institutions recommended that the order of documents within grant applications could also be reorganized so that the specific aims and research proposal would appear at the beginning of the application. Placing the budget, resources, and biosketches at the end would emphasize the increased importance of the science, rigor of the methodology, and innovation.
- Some societies/institutions noted that NIH should make efforts to pilot the proposed framework before implementation.

Some societies and academic institutions noted that NIH should conduct impact evaluations on whether the proposed framework post-implementation led to the expected outcomes (reducing bias, reviewer burden).

Responsive Societies and Academic Institutions

Societies	Academic Institutions
American Academy of Nursing	Boston University School of Medicine
American Association for the Study of Liver Diseases	Bowling Green State University
American Association of Colleges of Osteopathic Medicine	Brigham and Women's Hospital
American College of Emergency Physicians (ACEP)	Mass Eye and Ear/Mass General Brigham

American Epilepsy Society	Mass General Hospital
American Geriatrics Society	McLean Hospital
American Osteopathic Association	Mass General Brigham
American Physiological Society	Clemson University
American Podiatric Medical Association	Cold Spring Harbor Laboratory
American Society for Biochemistry and Molecular Biology	Florida State University
American Society for Cell Biology	George Mason University
American Society for Pharmacology and Experimental Therapeutics	Georgia Institute of Technology
American Society of Hematology	Harvard T.H. Chan School of Public Health
American Statistical Association	MIT
American Thoracic Society	Northern Illinois University
Association of Academic Physiatrists	Research Oakland University
Association of American Medical Colleges	Rowan University
Coalition for the Life Sciences (CLS)	Spaulding Rehabilitation Hospital
Council for the Advancement of Nursing Science (CANS)	Stanford University
Endocrine Society	Syracuse University
Federation of American Societies for Experimental Biology (FASEB)	The Edward Via College of Osteopathic Medicine
Federation of Associations in Behavioral and Brain Sciences	University of California, Merced
Good Science Project	University of California, Office of the President
Grant Professionals Association	University of Colorado - Anschutz Medical Campus
People for the Ethical Treatment of Animals	University of Hawaii at Manoa
Society for Academic Emergency Medicine (SAEM)	University of Nebraska at Omaha
Society for Pediatric Research	University of Pittsburgh
Society of Gynecologic Oncology	University of Wisconsin-Madison
The American Association of Immunologists	
The Humane Society of the United States	

Acknowledgments

The analysis was conducted by the Public Health Analyst Team at the NIH Center for Scientific Review.