REPORT OF THE COMMITTEE

ON

RATING OF GRANT APPLICATIONS

Table of Contents

May 17, 1996 (Revised)

Executive Summary

This report is from the Rating of Grant Applications (RGA) subcommittee of the Committee on Improving Peer Review (IPR). The subcommittee (referred to hereafter as "the Committee") was formed in the Fall of 1994, and was charged with examining the process by which scientific review groups rate grant applications and making recommendations to improve that process. Moreover, the Committee's task was to make its recommendations in light of what is currently understood regarding the science underlying psychological measurement and decision-making. The committee operated from the starting point that the rating system currently in use by NIH scientific review groups works reasonably well: no one appears to believe that poor quality science is consistently being given good scores nor that exceptionally good science is consistently being given poor scores. Nonetheless, in today's funding environment, it becomes increasingly important to ensure that scores are as reliable as they can be, and that program staff have the maximal amount of useful information on which to base their funding decisions.

The issues driving the current effort to improve the rating process include the tendency toward compression of priority scores at the better end of the scale, the generation and use of scores with the implication that they have more reliability and precision than they likely do have, weaknesses in current procedures for percentiling of priority scores, and the variable application of review criteria across committees and a sense of increasing focus on technical details rather than on the broader importance or potential impact of proposed research.

Committee discussions resulted in the defining of characteristics that should be found in any rating system used in peer review of research project grant applications. These characteristics, listed below, served as points of departure in subsequent discussions and ultimately in the development of the recommendations.

1) The rating assigned to an application should be a quantitative representation of scientific merit and should not represent any other property of the application.

2) The criteria used in the review of applications should include all aspects of the concept "scientific merit," and nothing else. Moreover, the criteria should be made as salient as possible to the reviewers and should form the only basis of both the quantitative ratings and the narrative critique of each application.

3) The ratings of all reviewers participating in a review should have the opportunity to be of equal influence in determining the final score of scientific merit for a given application.

4) The potential for "gaming" the system (i.e., consciously or unconsciously introducing inequities in the system based on factors other than scientific merit or distorting the assigned values representing scientific merit) should be minimized.

5) Results should be reported in a manner that is appropriate to summarize the totality of information contained in the ratings given by the review group as a whole. The numbers reported should neither overstate that information nor understate it.

6) Results should be reported in a form that is useful to institute/center officials in making funding decisions, and also in a manner that is informative to advisory councils/boards and applicants.

7) The rating system should encourage reviewers to make as fine discriminations as they are reliably able to make, especially in regions of the scale that might be near the funding lines of the Institutes and Centers.

8) Procedures should minimize the burden to reviewers prior to and at the review meeting.

9) Federal policy issues (e.g., gender/minority representation, protection of human/animal subjects) must be addressed appropriately.

The approach that the Committee used was to consult behavioral/psychometric experts and literature with regard to decision-making and evaluative processes. Next, extant data from previous peer review cycles were analyzed when the Committee believed those could illuminate some aspect of the Committee's deliberations, and simulations were developed to help explore the implications of the concepts being formulated. The Committee attempted to probe the ramifications of each approach being considered.

The major issues addressed by the Committee were: the review criteria and how they are used by the reviewers, the scale on which reviewers make their quantitative ratings, statistical manipulation of reviewer ratings to derive a score that would maximize comparability of scores across reviewers and review groups, and the necessity of pilot testing all new procedures before they are adopted or rejected.

The principal recommendations of the Committee follow:

1) *Recommendations Related to the Review Criteria*

Recommendation 1: The proposed, reformulated review criteria should be adopted for unsolicited research project grant applications. The reformulated criteria are:

<u>Significance</u>: The extent to which the project, if successfully carried out, will make an original and important contribution to biomedical and/or behavioral science.

Approach: *The extent to which the conceptual framework, design (including, as applicable, the selection of appropriate subject populations or animal models), methods, and analyses are properly developed, well-integrated, and appropriate to the aims of the project*

Feasibility: *The likelihood that the proposed work can be accomplished by the investigators, given their documented experience and expertise, past progress, preliminary data, requested and available resources, institutional commitment, and (if appropriate) documented access to special reagents or technologies and adequacy of plans for the recruitment and retention of subjects.*

Recommendation 2: Reviews should be conducted criterion by criterion, and the reviewers' written critiques should address each criterion separately.

Recommendation 3: Applications should receive a separate numerical rating on each criterion.

Recommendation 4: Reviewers should not make global ratings of scientific merit.

2) *Recommendations Related to the Rating Scale*

Recommendation 5: The rating scale should be defined so that larger scale values represent greater degrees of the characteristic being rated and the smaller values represent smaller degrees.

Recommendation 6: The number of scale positions should be commensurate with the number of discriminations that reviewers can reliably make in the characteristic being rated. An eight-step scale (0-7) is recommended on the basis of the psychometric literature; however, a maximum of 11 steps (0-10) are acceptable.

Recommendation 7: The rating scale should be anchored only at the ends. The performance of end-anchors should be evaluated and other approaches to anchoring should be investigated as needed.

3) *Recommendations Related to the Calculation, Standardization, and Reporting of Scores*

Recommendation 8: Scores should be standardized on each criterion within reviewer and then averaged across reviewers. The exact parameters for this standardization should be defined by an appropriately constituted group.

Recommendation 9: Scores should be reported on the eight-point scale used by reviewers in making the original ratings. Scores should be reported with an implied precision commensurate with the information contained in the scores. Two significant digits are recommended.

Recommendation 10: If a single score is required that represents overall merit, it should be computed from the three criterion scores using an algorithm that is common to all applications. The Committee favors the arithmetic average of the three scores: however, an appropriately constituted group should test and choose the algorithm to be used.

These recommendations cover a very broad scope of the initial review process, and any implementation of them must be accomplished in a way that will be minimally disruptive to the peer review system as it currently exists. In fact, implementation of some of them can proceed independently of the others. Ultimately, some may be implemented and others may not. Therefore, the Committee wished to comment on the way in which it sees the recommendations relating to each other.

1) The revised review criteria could be implemented without implementing any other recommendation. If they are implemented, they could be used with or without implementing the procedures for using them and for rating by criterion. Critiques could be written by criterion without rating by criterion.

2) The rating scale could be changed without changing anything else.

3) The analysis and recommendations in Section 5 (Summarizing and Reporting Ratings) could be applied to any scale measuring any aspect of scientific merit. The last two recommendations apply only if applications are rated by criterion.

Each of the recommendations flows from a basis in the literature of psychological and/or statistical measurement. Although the Committee made some effort to pilot test or simulate the systems it is recommending, every suggested change should be subjected to additional scrutiny and comment by all stakeholders in the review process and many should be pilot tested. The Committee believes that the details of implementation must be left to subsequent groups after this report has been circulated and subjected to comment, and after appropriate pilot testing and evaluation have been completed.

May 17, 1996 (Revised)

REPORT OF THE COMMITTEE ON RATING OF GRANT APPLICATIONS (RGA)

This committee is a subcommittee of the Committee on Improving Peer Review (IPR), chaired by Dr. Constance Atwell. The IPR Committee is charged with reviewing all aspects of the NIH peer review system and recommending changes that would result in improvements and/or streamlining of that system. During the course of IPR's discussions, a number of issues arose that appeared to be related to how reviewers assign ratings of scientific merit (priority scores) to applications. They were laid aside temporarily as being of a specialized nature, and a commitment was made to consider them at a later time. In the Fall of 1994, the Committee on Rating of Grant Applications (RGA) was formed to examine issues pertaining to this general topic. A list of the Committee membership is attached to this report. In addition, the Committee enlisted the aid of several experts in the fields of psychometrics and decision-making (list attached). The Committee is highly indebted to them for sharing their knowledge with us. Any inaccuracies of fact or interpretation in this report are the fault of the Committee and not these consultants.

1. Background and Events Leading to the Formation of the Committee

It is generally accepted that the rating system currently in use by NIH scientific review groups works reasonably well: no one appears to believe that poor quality science is consistently being given good scores nor that exceptionally good science is consistently being given poor scores. Nonetheless, because of the relatively small percentages of applications being funded by the Institutes and Centers (ICs) of NIH, it becomes increasingly important to ensure that scores are as reliable as they can be, and that program staff have the maximal amount of useful information on which to base their funding decisions. A number of review issues have become apparent, exacerbated by tightening constraints on available funds for the support of grants, and it was felt that many of these could be addressed by a thorough reexamination of review and scoring procedures at NIH. Among the issues that led to the formation of this Committee are the following.

- There appears to be an unrelenting tendency for priority scores to move toward the better end of the scale and hence to cluster in the "outstanding" region. Although it is arguable that applications are getting better and most science now being received by NIH is either excellent or outstanding, this compression of scores makes discrimination of the best science difficult. Furthermore, such compression inevitably leads to misleading or possibly even inaccurate distinctions when the scores from three consecutive meetings are combined in the process of percentiling. For example, ties at one priority score or large clusters of scores over a narrow priority score range create what might appear to be significantly large albeit meaningless percentile gaps at the next sequential priority score. In addition, random reversals can occur for applications from different review rounds and

this can lead to interpretations of relative merit that are spurious, thus leading to the possibility of unintended deviations from normal priority order when making funding decisions, particularly at the "payline."

- Scores are generated (by reviewers), calculated (by computer algorithms) and used (by program staff) as if they represented a higher degree of reliability and precision than they actually have. Setting aside triage/streamlining to simplify discussion, the current system has reviewers score applications on a 41-point scale (1.0 to 5.0 in 0.1 unit increments). These are clearly more scale positions than can reliably be discerned. The committee average ratings are then converted to and reported out as "priority scores" on a 401-point scale (100-500-- again, triage/streamlining not taken into consideration) and then converted to percentiles, which are reported out on a nominal 1000-point scale (0.0 to 100.0). Program staff are then put in a position of and held hostage to making and defending funding decisions based on these small and likely meaningless mathematical differences.

- The manner of percentiling *ad hoc* reviews is not satisfactory. Current percentiling procedures are based on the existence and continuity of chartered study sections or their equivalent. For *ad hoc* reviews, some other base, such as the combined voting patterns of all DRG committees, is used as a reference. The percentile thus derived has little defensible basis.

- There is a lack of salience of the criteria used by reviewers, potentially leading to the use of private and/or idiosyncratic review criteria within individual committees. In this regard, the General Accounting Office (GAO) issued a report in 1994 on the peer review systems used by the NIH, the National Science Foundation (NSF), and the National Endowment for the Humanities (NEH). The report was entitled "Peer Review: Reforms Needed to Ensure Fairness in Federal Agency Grant Selection." In it, NIH scientific review groups were criticized for the use of "unwritten criteria" in their evaluations of applications, most notably the requirement for preliminary data. The GAO concluded that the use of other than published review criteria leads to a lack of fairness and a possible advantage for the "insider" or reviewer in writing applications to address the criteria actually used. On another level, there is growing concern that critiques focus more on technical details of the proposed research rather than on the broader importance or potential impact of the project. Thus the criteria as currently defined may not serve as well as they could.

- The current system of voting priority scores originated at a time when a majority of submitted applications were funded; nonetheless, it continues to be used with most success rates now in the range of 10 to 25%.

The overall perception of the current rating system is that it has developed over many years in an evolutionary way, responding to various pressures and issues that may or may not be relevant to

current conditions. It now seems appropriate to take an in-depth look at the evaluation and scoring process in light of what is currently known about psychometrics and decision science and to bring this scientific information and knowledge to bear on the review of grant applications at NIH.

Psychometric principles were first explicitly applied to the grant review process within the Public Health Service in 1989 when the National Institute of Mental Health, as a component of the Alcohol, Drug Abuse, and Mental Health Administration, convened a workgroup of scientists to address priority score descriptors for use in scoring applications. The group was made up of non-Federal scientists with expertise in psychometrics, evaluation, and decision research. The charge to the group was to establish a descriptor scale that was psychometrically valid and that would facilitate the spreading of priority scores over the entire numerical scale. They met the charge, developing descriptors based on results of questionnaires and empirical examination of the question. The group did not stop there, however, and provided advice and comment on other aspects of scoring that provided some of the seeds that have germinated in the current committee. Several members of that workgroup as well as other non-Federal scientists have served as consultants to RGA.

1.1. Charge to the RGA Committee

The RGA Committee was charged with examining the process by which scientific review groups rate grant applications and making recommendations to improve that process. Moreover, the Committee's task was to make its recommendations in light of contemporary thinking in the behavioral sciences as it relates to psychological measurement and decision-making. Throughout the Committee's deliberations, it focused on procedures used to review research project grant applications, specifically R01s submitted as investigator-initiated projects.

It should be emphasized that the Committee did not have the specific objective of streamlining the review process, but rather of making it better. However, the Committee was mindful of the workload implications of all the options it discussed, and it tried to ensure that there was a value-added potential in each recommendation it made.

1.2. Basic Operating Framework

1.2.1. Functions of Peer Review

In defining the scope of its activities, the Committee viewed the initial (scientific) review of applications as serving two functions:

- To assess the "scientific and technical merit" of an application (as referred to in the PHS Act) through a narrative and one or more quantitative indices (scores). "Scientific and technical merit" will be referred to simply as "scientific merit" in this report.

- To comment on other aspects of the application (e.g., biosafety issues, human subject and animal welfare concerns, budgetary considerations) that fall outside the concept of scientific merit per se. Such comments should be clearly separated from the discussion and rating of scientific merit.

The Committee interpreted its charge to pertain only to the review of applications for scientific merit, but it wishes to emphasize that the review process must be carefully structured to separate the determination and reporting of scientific merit from all other advice that scientific review groups might be asked to give or that they might wish to give. Only by doing this will reviewers give judgments of scientific merit that are as uncontaminated as possible by their views of issues not properly part of scientific merit.

1.2.2. Characterization of an Ideal Rating System

During the course of its discussions, the RGA Committee developed a set of "Guiding Principles" which then served as points of departure for developing its recommendations.

The rating system used in peer review of research project grant applications should have the following characteristics:

1) The rating assigned to an application should be a quantitative representation of scientific merit and should not represent any other property of the application.

2) The criteria used in the review of applications should include all aspects of the concept "scientific merit," and nothing else. Moreover, the criteria should be made as salient as possible to the reviewers and should form the

only basis of both the quantitative ratings and the narrative critique of each application.

3) The ratings of all reviewers participating in a review should have the opportunity to be of equal influence in determining the final score of scientific merit for a given application.

4) The potential for "gaming" the system (i.e., consciously or unconsciously introducing inequities in the system based on factors other than scientific merit or distorting the assigned values representing scientific merit) should be minimized.

5) Results should be reported in a manner that is appropriate to summarize the totality of information contained in the ratings given by the review group as a whole. The numbers reported should neither overstate that information nor understate it.

6) Results should be reported in a form that is useful to institute/center officials in making funding decisions, and also in a manner that is informative to advisory councils/boards and applicants.

7) The rating system should encourage reviewers to make as fine discriminations as they are reliably able to make, especially in regions of the scale that might be near the funding lines of the Institutes and Centers.

8) Procedures should minimize the burden to reviewers prior to and at the review meeting.

9) Federal policy issues (e.g., gender/minority representation, protection of human/animal subjects) must be addressed appropriately.

The Committee agreed that it should not be constrained by current or familiar practice in and of itself, but should be prepared to propose any workable system, even a radically different one, if such a system should appear to be clearly superior on the basis of available theory and evidence. The approach that the Committee used in attempting to design a system that would meet the above requirements was to consult the behavioral/psychometric experts and literature with regard to decision-making and evaluative processes. Next, extant data from previous peer review cycles were analyzed when the Committee believed they could illuminate some aspect of the Committee's deliberations, and simulations were developed to help explore the implications of the concepts being formulated. The Committee attempted to probe the ramifications of each approach being considered.

1.2.3. Issues Addressed by the Committee

The following major issues emerged as foci for the Committee:

- The review criteria and how they are used by the reviewers (Section 3).
- The scale on which reviewers make their quantitative ratings (Section 4).
- Statistical manipulation of reviewer ratings to derive a score that would maximize comparability of scores across reviewers and review groups (Section 5).
- The necessity of pilot testing all new procedures before they are adopted or rejected.

2. Maximizing the Quality of Scientific Merit Ratings

The task of rating a grant application for perceived scientific merit can be conceptualized as a complex mental activity which must be structured in a way that allows reviewers to produce ratings that faithfully represent their best judgments of the application relative to the properties being rated. Unfortunately, the quality of ratings is difficult to evaluate because there is no "gold standard" of scientific merit to which the ratings of any given reviewer or reviewers can be compared. (Such measures as citation counts or post hoc judgments of scientific impact are only correlates of scientific merit and generally cannot be measured until much later than when the ratings of applications are made.) The task of the Committee, then, was to assess indirectly the quality of current and proposed procedures in any way possible. One approach was to look for parallels in the psychological research literature that would suggest that one procedure might be better than another. Another was to appeal to "face validity" whenever possible--that is, to seek the judgments of individuals thoroughly familiar with all aspects of the review task. Extant empirical data were analyzed and simulations developed whenever they could indirectly illuminate an issue.

The Committee noted that behavioral scientists conceptualize the rating of complex stimuli (such as applications) in different ways, emerging from different traditions of behavioral research.

One way of conceptualizing the rating task flows from the large body of psychological research on the manner in which people estimate the magnitude of a physical characteristic of a stimulus (for example, the brightness of a light, the pitch of a tone, or the loudness of a sound). Extrapolating from this body of research, one tends to conceptualize the rating of applications as one of magnitude estimation in which a reviewer is asked to estimate the magnitude of the property "scientific merit" inherent in a given application. Under the simplest version of this view,

scientific merit is measured as a single variable, the magnitude of which can be perceived and estimated by certain experts (reviewers), and review criteria form a working definition of the concept of scientific merit to be used while estimating the degree of scientific merit in an application. Ultimately, it is left to the reviewers to estimate the variable using the totality of their own knowledge, experience, wisdom, and understanding of the concept "scientific merit."

Another, more recent, area of behavioral science that is relevant to the present task is the study of how people make complex decisions. The decision scientist sees the rating task as one in which each reviewer makes a complex decision about assigning an application to a position on a scale of scientific merit. The assigned rating is seen as being the end product of a number of component decisions. Within this framework, the way to improve the quality of the decisions is to structure the overall task so that the reviewer can perform component tasks that are relatively easy and can avoid tasks that are difficult and performed less reliably. In particular, a significant body of literature (e.g., Dawes, Faust, and Meehl (1989), Knaus, Wagner, and Lynn (1991), Zuckerman (1976)) indicates that, in most situations, people are relatively poor at reliably making complex global judgments (sometimes called "clinical" judgments) in which they are required to integrate information from several sources or from separate component decisions. On the other hand, people are relatively better at making judgments on the primary components of a more complex task.

After extensive discussion, the Committee combined both approaches in its conceptualization of the rating task. It concluded that a way to improve assessments of scientific merit is to ask reviewers to make their ratings criterion by criterion so that ratings on each criterion are relatively simpler magnitude estimations of the primary dimensions of "scientific merit." Evaluations of an application on each of the individual review criteria, taken together, represent a *disaggregated* evaluation of overall scientific merit, as contrasted with the single global judgment that has traditionally been provided by reviewers.

For the rating of each criterion, the Committee used traditional psychometrics to develop the scales on which ratings would be made.

3. Review Criteria

If scientific merit is viewed as a complex concept composed of a number of components, each represented by a review criterion, then the first step in developing a system for evaluating scientific merit is to develop review criteria which, taken together, fully comprise scientific merit and can be individually and independently evaluated by reviewers.

3.1. Current Review Criteria

The current review criteria for evaluating research project grant applications are stated as follows:

- scientific and technical, or medical, significance and originality of the goals of the proposed research;
- appropriateness and adequacy of the experimental approach and methodology proposed to carry out the research;
- qualifications and research experience of the principal investigator and staff, particularly but not exclusively in the area of the proposed research;
- availability of resources necessary to conduct the research;
- the appropriateness of the proposed budget and duration relative to the proposed research; and
- adequacy of plans to include both genders and minorities and their subgroups as appropriate for the scientific goals of the research. Plans for the recruitment and retention of subjects will also be evaluated.

The Committee agreed that the review criteria currently in use include the principal aspects of scientific merit. However, it was also generally agreed that they could be stated more succinctly and in a way that would better highlight and focus reviewers' attention on the dimensions underlying merit. The goal was therefore to construct a system that would explicitly address the primary dimensions of scientific merit: significance, research approach, and feasibility.

3.2. Recommended Review Criteria

Significance: The extent to which the project, if successfully carried out, will make an original and important contribution to biomedical and/or behavioral science.

Approach: The extent to which the conceptual framework, design (including, as applicable, the selection of appropriate subject populations or animal models), methods, and analyses are properly developed, well-integrated, and appropriate to the aims of the project

Feasibility: The likelihood that the proposed work can be accomplished by the investigators, given their documented experience and expertise, past progress, preliminary data, requested and available resources, institutional commitment, and (if appropriate) documented access to special reagents or technologies and adequacy of plans for the recruitment and retention of subjects.

The following issues are to be considered and commented upon by the scientific review group as appropriate but are not to enter into the assessment of scientific merit.

- where activities are involved that could have an adverse effect on humans, animals, or the environment, the adequacy of the proposed means for protecting against or minimizing such effects.

- issues pertaining to applications from foreign institutions.

3.3. Other Issues Relating to Review Criteria

Other Issues Considered by the Committee in developing the above criteria:

3.3.1. Possible Additional Criteria

3.3.1.1. "Best Buy"

At various times in the early 1990s, it has been proposed that initial review groups should provide assessments of the cost-effectiveness of proposed budgets and that cost-effectiveness should be part of scientific merit. In 1994, a study was conducted of (1) the ability of reviewers to make such assessments of every application they reviewed and (2) the usefulness of such information to program officers in making funding decisions. Results showed that reviewers did not make cost-effectiveness assessments and comments that were useful to program officers in most cases. (*NIH Experiment to Augment Budget Evaluations in Scientific Review.* Working Group on Cost Management. Michael Goldrich, Chair)

The Committee considered it useful for the reviewers occasionally to note that a particular project was especially cost-effective, but there was no enthusiasm for adding this as a mandated separate criterion to be included in the concept of "scientific merit."

However, reviewers should be encouraged to make note of any unusually high or low cost-effectiveness either in their evaluations of requested resources or in a separate administrative note.

3.3.1.2. Appropriateness of Investment

The Committee on Interactions with Howard Hughes Medical Institute Supported Scientists suggested that the Committee include a criterion based on the total funding within the applicant laboratory and the likelihood that the proposed project would proceed with or without the support requested in the current application. That committee noted that reviewers may want to give their opinion on whether Federal dollars should be spent on the project, regardless of its scientific merit. The RGA Committee did not think that this was an appropriate function of review committees and certainly was not a part of the concept "scientific merit." Rather, the Scientific Review Administrator should discourage discussion or voting on issues other than the stated review criteria. Discussion of the resources available to the project is appropriate under the Feasibility criterion, however. Administrative notes may be used to convey additional information to program staff that may be helpful in their consideration of applications for funding, even though they are not relevant to the scientific evaluation of the project.

3.3.1.3. A Creativity/Innovation Criterion

A suggestion was made that creativity and innovation should be accorded the status of a separate, fourth criterion. The Committee does not agree with this view for the following reasons:

1) Creativity/innovation is currently included in the Significance criterion in the phrase "...*original* and important contribution... (emphasis added)."

2) Creativity/innovation, in and of itself, is not necessarily a hallmark of scientific merit. It must be coupled with excellence on the other dimensions of scientific merit for it to increase an application's appeal as a candidate for the use of public funds.

3) By definition it is a relatively rare trait and thus is best assessed and described in the narrative for each criterion as appropriate.

3.3.2. Relations Among Proposed Criteria

3.3.2.1. Implicit Hierarchy

As noted, the intention of the Committee was to create review criteria that would be conceptually independent of each other but inclusive when taken together. There was considerable discussion, however, about the relationship among the proposed criteria. For example, there was discussion about the extent to which a low assessment on any one criterion would obviate the need for extended discussion of the application on the other criteria. It was agreed that the Significance of a project is paramount, followed by Approach and then Feasibility; however, the Committee agreed that all three criteria should be weighted equally for the purpose of rating and calculation of overall scores (*vide infra*).

3.3.2.2. Implications for the Review Process

The Committee recommends that the review be conducted by considering all reviewers' evaluations of Significance first, then Approach, and finally, Feasibility. (Although a majority of the Committee favors a procedure in which applications would be discussed and rated criterion by criterion, the procedure by which the criteria are utilized in review sessions should be discussed and pilot tested in a variety of scientific review groups before a single procedure is adopted.) Applications that did not fare well on Significance would receive abbreviated discussions of their Approach and Feasibility.

Recommendation 1: The proposed, reformulated review criteria should be adopted for unsolicited research project grant applications.

Recommendation 2: Reviews should be conducted criterion by criterion, and the reviewers' written critiques should address each criterion separately.

Recommendation 3: Applications should receive a separate numerical rating on each criterion.

3.3.2.3. A Global Rating of Scientific Merit

In the current procedure for reviewing applications, each application that is discussed by the IRG is assigned a priority score rating by each reviewer. This rating represents an overall, global rating of the application's scientific merit taking into consideration the entire set of review criteria. In the recommended procedure, reviewers would rate each application separately for each criterion. The question then arises regarding whether each application should have a global rating of scientific merit associated with it, and, if so, how that rating should be derived.

The committee does not believe that a global score of merit is essential to characterize an application; it is comfortable with a profile of three scores in that regard. Nevertheless, it may be that the Institutes and Centers require a single, global score of merit to develop funding queues and determine budget allocations to their various programs. If it is determined that a global score is required, there are two approaches that could be taken to develop it.

One way of developing a global score of scientific merit would be to ask reviewers to make ratings of overall merit, much as they do now. Under this model, reviewers would make ratings by criterion and then make a final rating of overall scientific merit. The Committee discussed this option with its consultants and found that there was considerable disagreement on this issue.

Some consultants and Committee members preferred a global rating because it allowed reviewers flexibility in weighting the criteria. These individuals argued that ratings of overall merit more fully reflected reviewers' judgments about the merits of each application, and that expert committees are in the most knowledgeable position to rate global scientific merit.

Other consultants argued that research in a variety of settings involving decision-making behavior showed that global ratings of complex stimuli are less valid (when compared to a gold standard) than are ratings that are derived algorithmically from ratings of

components of the stimuli. They also point to the curious phenomenon, documented in Dawes, Faust, and Meehl (1989), that although global ratings tend to be less valid than derived ratings, raters commonly believe that their global ratings are more valid than any more "objective" measures.

The Committee agrees substantially, though not unanimously, with the latter view and recommends against having reviewers give global ratings. The following reasons are the basis for that recommendation:

1) Such ratings are likely to be less valid than scores computed from criterion ratings.

2) If reviewers are allowed to make global ratings, there may be a tendency for them to give the criterion ratings less attention and, in fact, to subconsciously manipulate the criterion ratings to "fit" the global rating. Therefore, not only is a global score computed from criterion ratings likely to be more valid than one based on global ratings, but also the act of making global ratings can lower the quality or credibility of the criterion ratings.

3) Global ratings based on the individual reviewer's judgment of what constitutes good science has the potential of incorporating criteria that are not explicitly defined and accepted and hence may not be appropriate.

Recommendation 4: Reviewers should not make global ratings of scientific merit.

4. The Rating Scale

4.1. Principles of Scale Development

The Committee agreed upon two principles to guide its deliberations in this area:

1) There is a distinction between the scale on which reviewers are asked to make their ratings and the way in which their ratings are summarized and reported. In the current system, reviewers make their ratings on a 41-point scale ranging from 1.0 to 5.0, and their ratings are summarized by taking the mean of the reviewers'

ratings and multiplying it by 100. This report will discuss these two topics separately, because quite different considerations must be brought to bear on each. In order to maintain this distinction, the term "rating" will be used to denote the quantitative judgment made by a reviewer, and the term "score" will be used to denote a number that summarizes and reports the ratings of a set of reviewers judging a given application. (Thus, in the current system, ratings are made by reviewers on a scale from 1.0 to 5.0 and priority scores and percentile ranks are the scores that are computed from those ratings to summarize and report them.) The present section speaks only to the scale on which reviewers make their ratings.

2) The rating scale to be recommended will be consistent with how people make judgments about complex stimuli. It is intended neither to over-estimate nor under-estimate reviewers' abilities to make reliable and (presumably) valid judgments about the scientific merit of each application. (It must be borne in mind, however, that there is no "gold standard" measure of scientific merit available that can be used to assess the reliability and validity of any rating process. Thus, the Committee was forced to extrapolate from the relevant psychological literature and also bring to bear the pragmatic lessons of 50 years of peer review at NIH.)

4.2. Number of Positions and Polarity

It is generally accepted that higher scale positions reflect higher values of the item being rated. Nonetheless, it has been standard practice at NIH to invert the scale, using 1 as best and 5 as worst on the current 5-point scale. The Committee notes that the scale for rating grant applications at NIH should follow the more widely accepted convention.

Recommendation 5: The rating scale should be defined so that larger scale values represent greater degrees of the characteristic being rated and the smaller values represent smaller degrees.

The RGA Committee has consulted with experts in the scientific community regarding the number of increments on the rating scale. Balance is needed between providing sufficient increments to allow raters the maximum discriminable judgment and providing excessive numerical refinement such that differences between adjacent increments are not reliably different. There is clear agreement that the 41 points in the current scale are excessive. Surveys of literature relating to the development of rating scales indicate that raters in a task such as this can, at a minimum, make seven discriminations validly and reliably, and thus could use a seven-step scale (c.f., Miller, 1956; Landy and Farr, 1980, Cicchetti, Showalter, and Tyrer; 1985). This literature also states that up to eleven steps can be used without loss of reliability.

The Committee favors the use of a seven-step scale, but recognizes that there may be advantages to a ten-step scale. The chief advantage of a 1-7 scale is that it is clearly within the ability of reviewers to use effectively. Also, it may induce reviewers to use the entire scale in order to make the discriminations that they want to make. A 1-10 scale has the advantage of familiarity for reviewers, since they use a decimal scale in many numeric applications of their daily lives; however, its additional scale positions may invite reviewers to use only the upper part of the scale (as they do currently).

The Committee discussed whether or not the rating scale should include a zero at the low end or should be anchored on the low end by a one. Although either can be accommodated arithmetically, the Committee saw an advantage to offering reviewers a scale position that indicated an absence of merit or a judgment of "unacceptable" relative to the criterion being rated. Therefore, the Committee favors the rating scale having a low position anchored by the digit "0" rather than "1."

Recommendation 6: The number of scale positions should be commensurate with the number of discriminations that reviewers can reliably make in the characteristic being rated. An eight-step scale (0-7) is recommended on the basis of the psychometric literature; however, a maximum of 11 steps (0-10) are acceptable.

4.3. Comparability of Ratings

Because scores of scientific merit are used as a primary factor in funding decisions, it is essential that they be comparable with each other regardless of whether they derive from reviews held at different times and/or in different review groups. This requirement is very important in determining the instructions given to reviewers and in the way ratings are treated statistically in deriving scores to be reported.

4.3.1. Rating versus Ranking

There are two basically different ways that reviewers can be instructed. First, they can be instructed to consider each application without reference to other applications, to compare it to the set of criteria, and to rate it on the degree to which it meets the criteria. For example, reviewers could be asked to rate the Significance of an application in an absolute sense, without reference to the other applications that are before the group.

Alternatively, reviewers could be asked to arrange the set of applications with which they are dealing so that each successive application has more of a given property (e.g., Significance) than the one before it. This would be an example of a ranking approach to determining scientific merit.

Throughout the history of NIH peer review, reviewers have been instructed to rate applications relative to the review criteria and without reference to other applications. This has been the procedure despite the general finding that people are much better at comparing stimuli currently before them than they are at comparing a stimulus with an abstract standard. The principal reason why a rating approach has been used is that scores of scientific merit must be as independent as possible from the specific contexts in which they were reviewed, because they will usually be placed in funding queues with other applications reviewed by other review groups or by the same group at a different meeting.

The comparability of ratings across reviewers (and review groups) requires that reviewers use the rating scales in the same way to the greatest extent possible. This is referred to as the "calibration problem." In the rating task, it is very difficult to anchor the various scale positions so that all reviewers are calibrated in the same way and so that a given numerical rating given by different reviewers will represent the same cognitive appraisal.

4.3.2. Criterion Referencing versus Norm Referencing

Generally speaking, there are two ways to "anchor" the various positions on a scale such that different judges (reviewers) will use them in the same way--norm referencing and criterion referencing. In norm referencing, scale positions are anchored by referring them to percentages of a reference population of stimuli. (For example, judges could be instructed to use an eight-point scale such that equal numbers of some population of stimuli would end up in each of the scale positions. In such a case, each scale position would be anchored to a particular segment of the distribution of stimulus values.) In criterion referencing, scale positions are anchored by referring them to stimuli that are defined as being instances of those positions, and/or by referring them to verbal descriptions of the stimuli that should be placed in each scale position.

In the current system, both types of referencing are used to some extent. Criterion referencing is used when regions of the scale are anchored by referring them to the meanings of common adjectives, as follows:

Outstanding	Excellent	Very Good	Good	Acceptable
1.0 - 1.5	1.5 - 2.0	2.0 - 2.5	2.5 - 3.5	3.5 - 5.0

Although these anchors are frequently used by reviewers in communicating their assessments to other reviewers, their use has not prevented groups as a whole from rating larger and larger percentages of applications in the "Outstanding" and "Excellent" ranges. Also, many experienced reviewers appear to use the adjectives merely as labels for regions on the scale and not according to their normal English meanings.

An alternative view of these adjectival anchors is that they actually have continued to provide appropriate anchors for the various regions of the scale but that the numbers of outstanding and excellent scores have, in fact, risen dramatically over recent years due to increasing sophistication in the community of investigators and the increasing proportion of amended applications. Although there is no easy empirical way to evaluate the actual anchoring power of the current set of adjectives, the Committee believes that both views are true--that adjectives are not highly effective anchors in the present system and that the average caliber of applications has risen and will continue to rise.

The Committee discussed the possibility of developing verbal descriptions of each scale position and then relying on them to mean the same thing to all reviewers and thus anchor the scale positions; however, this approach is not recommended at this time. Nevertheless, this approach might be helpful if developed through the use of empirical studies.

A form of norm referencing is currently used when reviewers are instructed to label approximately one-half of the applications they review as "unscored." (In this case, reviewers are told to imagine the reference population as being all applications that they would expect to come to that review group over several review cycles.)

To norm reference the rating scale in the present situation, the following instructions could be given to reviewers:

> *"Imagine the population of all applications that might be assigned to this particular review group over the long term. Place any application in the scale position '1' if it falls in the lowest one-seventh of the population on the criterion being rated. Similarly, the scale position '7' should be used*

to represent the highest one-seventh of the population. The scale positions '2' through '6' should then be used to represent approximately equal percentages of the reference population in a similar fashion."

Ultimately, the Committee is under no illusions about the power of either norm or criterion referencing to ensure that reviewers use scales comparably. Any sort of anchoring takes the form of instructions to reviewers -- instructions that may or may not achieve the objective of equal calibration.

Recommendation 7: The rating scale should be anchored only at the ends. The performance of end-anchors should be evaluated and other approaches to anchoring should be investigated as needed.

Another approach to ensuring comparability is *post hoc* standardization, where ratings are statistically processed to achieve comparability of scores. This subject will be treated in detail in Section 5.

5. Summarizing Ratings and Reporting Scores

As indicated in previous sections of this report, the Committee separated the rating of an application by reviewers on one or more scales related to scientific merit from the summarizing and reporting of those results. This section addresses the second issue.

5.1. Level of Precision

The Committee held extensive discussions about the level of precision to which scores should be reported. Some held the view that because reviewers are rating applications on a 0-7 scale using integers only, scores should be reported on the same scale, also using only integers, because to do anything more would be to exaggerate the precision of the data. Others pointed out that the purpose of a score is to summarize the information contained in all the ratings obtained from reviewers and, therefore, the amount of information to be represented by a score was greater than that in a single rating and thereby requires the use of more than one significant digit. The expert consultants suggested up to three significant digits; however, the Committee recommends two digits as a conservative representation of the information present in a set of ratings. The Committee acknowledged that often as few as two to five reviewers intensively study and critique a given application while the others may read the application less thoroughly, listen to the critiques, and participate in the ensuing discussion. Therefore, the number of truly independent, fully informed ratings may be somewhat fewer than the number of ratings actually given. There is no way to estimate this number, but it is probably no less

than the number of assigned reviewers/readers and certainly no more than the number of individuals rating the application.

The Committee heard a view that favored a "binning" approach to reporting scores--that is, reporting scores as falling into one or another of a very few "bins." The system used by NSF was considered where only five degrees of scientific merit are distinguished. Many observers of the NIH extramural system believe that binning would redress what is seen as an exaggerated dependence on the priority score or percentile rank in making funding decisions at the expense of portfolio balance, program priorities, and other factors that should enter into any funding decision.

The Committee believes that it is beyond its charge to recommend whether binning would be beneficial to program staff in making funding decisions. Rather, the Committee took the view that its charge is to recommend a system that would fairly and accurately summarize and report the information provided by reviewers. Therefore, the Committee takes no stand on whether the information content of scores should be reduced to accommodate procedures related to funding decisions, but it believes that to ask reviewers to use a scale with fewer than seven positions prohibits them from giving the NIH information that they can validly and reliably give.

5.2. Comparability of Scores Across Review Groups

There has been a long-standing concern about the comparability of scores across review groups. Most funding units in NIH consider applications from many different scientific review groups as competing for the same funds, and thus it is desirable that scores of scientific merit be directly comparable with each other, regardless of the scientific review group from which they come and regardless of the review cycle in which the application was reviewed.

5.2.1. Percentiling

For over a decade, the solution adopted by NIH to ensure the comparability of scores across groups has been to percentile scores within groups. The history of percentiling in DRG is described in the paper, *Percentiling of Priority Scores Assigned by NIH Initial Review Committees: Background and Specifications* (Information Systems Branch, Division of Research Grants, May 17, 1995). For standing initial review groups, after a given group has completed its work and the priority scores have been calculated, the scores are rank-ordered within all those reviewed by the given group for the present and previous two cycles of reviews. This rank-order is then converted to a percentile rank, with 0 being the best and 100 being the poorest. Reviews in ad hoc or other "non-qualifying" study sections

(i.e., those that have reviewed fewer than 25 applications over the past three cycles are percentiled against a pool of applications from all qualifying groups.

The pros and cons of percentiling by groups are relatively well known. The greatest advantage is that percentiling neutralizes any inflationary tendencies and also encourages reviewers to spread their scores across the full scale. Under percentiling, other factors being equal, approximately equal percentages of applications will be funded from each scientific review group, regardless of the absolute scores given to the applications. For precisely this reason, the greatest disadvantage of percentiling is that approximately equal percentages of applications from each group are funded, regardless of the perceived quality of science reviewed by the various scientific review groups. Thus, percentiling makes the tacit assumption that the overall quality of applications coming to all scientific review groups is about the same. To the extent that this is not the case, an "entitlement" to equal funding for each scientific review group is created that does not necessarily correspond to scientific merit.

A second disadvantage of percentiling is that it transforms the distribution of scores from whatever was given by the scientific review group to a rectilinear ("boxcar") shaped distribution. This has the unfortunate characteristic of appearing to spread out scores that were highly clustered in a narrow region. For example, if reviewers give many scores in the 130-135 range, the percentile for a priority score of 130 may differ markedly from the percentile for that of 135, disguising the fact that there may be no meaningful difference between the two scores.

A third difficulty with percentiling within groups is that there must be intact scientific review groups with histories for the system to work. To the extent that a significant proportion of reviews are done using more flexible special emphasis panels or special review groups where individual reviewers may attend different groups in different review cycles, there is no group within which to percentile.

The Committee searched for alternative ways of processing scores for comparability that would not have the disadvantages of percentiling within groups. After considerable discussion with its consultants, the Committee determined that no change in a rating system can completely solve the entitlement problem. One solution clearly beyond the scope of a rating/scoring system would involve some sort of comparative evaluation of the various disciplines of science made by a source external to the review process. However, a change in scoring procedures can serve to facilitate at least a partial remedy to entitlement. If a system were adopted that would "normalize" ratings within a given reviewer rather than within a review group as is now the case, then by mixing reviewers across groups, the

standards of scientific excellence of the groups would converge on each other and entitlement should diminish. Developing a system for assigning reviewers to review groups where mixing reviewers was the rule goes considerably beyond the scope of this committee's charge.

5.2.2. Standardizing

An alternative to percentiling would be to standardize the scores of each scientific review group, that is, to set the mean score for each scientific review group to a common, arbitrary value and, similarly, to set the standard deviation to a common value. Such a transformation eliminates differences in central tendency and in variability of scores across groups, but retains the shape of the distribution of scores originally created by the scientific review group, thus also retaining the relative differences among scores.

5.2.3. Transforming Scores by Group or by Individual

An alternative to adjusting scores within groups would be to adjust them within individual reviewers. Either percentiling or standardizing could be used within reviewer. The advantages of such a system would be that adjusting scores by group would be unnecessary, that reviewers could transport their individual voting histories from group to group without problem, and that review groups could be custom-tailored to sets of applications instead of forcing applications into pre-existing groups.

The above alternatives can be conceptualized in a four-fold table as follows:

Unit of Adjustment	Percentile	Standardize
By Group	**Current System**	Not Recommended [2]
By Reviewer	Possible [1]	**Recommended**

[1]One could percentile within reviewer much as the current system percentiles within group; then individually based percentiles could be averaged to create a global score for each application. This system would have the advantages of allowing flexible scientific review group membership, but would not retain the original shape of each reviewer's distribution of ratings.

[2]This strategy would not preserve the flexibility inherent in within-reviewer adjustment of scores.

The Committee recommends that the four options in the above table be intensively compared with respect to possible implementation. The Committee has not been able to conduct an analysis of the alternatives to the depth that it would prefer. The Committee favors the option of standardizing scores within reviewer as offering the greatest potential for eliminating unwanted differences in rating behaviors across individuals and groups while preserving the essential information contained in the ratings. This approach appears also to preserve flexibility in the assignment of reviewers to scientific review groups to meet the changing needs and trends in the science being reviewed.

Recommendation 8: Scores should be standardized on each criterion within reviewer and then averaged across reviewers. The exact parameters for this standardization should be defined by an appropriately constituted group.

5.2.4. Implications for Streamlined Review

The Committee wishes to point out that its recommendation to use standard scores assumes that all applications reviewed (or approved) are scored. Current DRG practice is to ask reviewers to identify the poorest 50 percent of the applications being reviewed so that they will not be discussed during the scientific review group meeting. Such applications are not rated and are formally designated only as "not scored." If this procedure is to be retained, the alternatives of standardizing and percentiling should be reevaluated taking into consideration the complexities introduced by having a significant proportion of applications unscored in each review.

The Committee recommends that the streamlined review procedures that have come into use in the last year or so be continued, but proposes several changes to: 1) respond to concerns by reviewers and the extramural community about the triage process, and 2) accommodate the needs of the proposed new scoring system.

The proposed change from the current "streamlined" review process would have each application considered to be non-competitive (or in the poorer half of those reviewed by a particular group) brought up very briefly at the meeting; and the assigned reviewers would address each of the review criteria in a sentence or two, highlighting the primary reasons for the application being considered in the lower half. The importance of this additional brief step is to: a) let all reviewers know what the primary reasons are for these assessments; b) allow for "discovered" disagreements to be resolved through discussion when it might affect the group recommendation; c) provide a sense that recommendations for all applications are

the recommendations of the entire review group; d) provide program staff with some information about the primary reasons for the recommendation; e) make it possible to prepare a brief summary of the review (to accompany the written critiques of the assignees) that would list the primary reasons for the recommendation for the benefit of the investigator/applicant, and f) make it possible for reviewers to assign scores to all applications so that applicants and staff will have some sense of where an application lies within the distribution.

Experience in NIMH, where a similar process has been used during the past year, has shown that the time saved at the review meeting is conserved, and that only rarely does it lead to extended discussions. When discussion does result from this process, it is in the best interest of the applicant and the NIH to have the issues aired. The benefits to the applicant and program staff listed above are reinforced by benefits related to implementing the proposed new rating system. If all applications are scored, even though some are not discussed in depth, it will be possible to get a better estimate of each reviewer's scoring behavior and standardizing within the reviewer will be more reliable.

What is preserved is the essence of streamlining, the principle that reviewers do not spend any more time on any application than is needed to make a recommendation. What is better about this approach is that all applications are considered by the full review group; the recommendation is more of a committee recommendation; and there is some sense that the committee members agree on the primary reasons for the recommendation.

5.2.5. A Reference Distribution for Standardization

Just as percentiling requires a reference distribution of scores against which to percentile any given score, so too does standardizing require that a reference distribution be defined against which any score is standardized. In particular, a standard score, *z*, is defined as: $z = \frac{(x - \overline{X})}{std.dev.}$, where *x* is any observed score from the reference distribution, $\overline{X}$ is the mean of that reference distribution, and *std.dev.* is the standard deviation of that distribution. DRG has defined the reference distribution for percentiling as the current round plus the previous two rounds of reviews done by a given group, with a minimum of 30 reviews having been done during that time. The Committee recommends that an essentially similar definition be used for standardizing within reviewer.

Although the Committee has not considered in detail how the reference distribution should be defined, the following is one possible definition that the

Committee believes would work: The reference distribution would be the reviews done by a given reviewer over the current and previous two review rounds with a minimum of 25 applications being scored by that reviewer over that time. For DRG-based reviews, certain reviews that the reviewer may have participated in would not be included, specifically, any reviews of applications submitted in response to RFAs and reviews of mechanisms other than R01s and R29s. The problem of how to deal with a new or occasional reviewer is a challenging one and is essentially the same problem as currently dealing with *ad hoc* or special review groups relative to percentiling. A number of approaches should be explored. For example, each new reviewer could be provided with a set of 25-30 artificially generated scores for each criterion representing an appropriately generic distribution. Each time the new reviewer rated an application, the actual rating could replace a rating from the generic distribution until the reference distribution included only the reviewer's ratings.

5.3. The Metric on Which Scores Will Be Reported

Standard scores have a mean of 0 and a standard deviation of 1. If scores of scientific merit were reported in those units, roughly half of the scores would be negative, and all would be small in absolute magnitude (few greater than 2). Thus, the Committee believes that the scores should be converted back to the original metric of the ratings, a scale of 0 to 7.

The Committee discussed the merits of reporting scores on a metric different from the one on which ratings are made. For example, scores could be reported on a scale from 0-100 with mean of 50 and a standard deviation of 15. The advantage of such an arbitrary metric is that scores would never be confused with ratings. (Similarly, in the current system, ratings are never confused with priority scores, and neither of these are confused with percentiles because they are on different scales. On the other hand, percentiles are often confused with success rates, since both are on the same scale.) The majority of the Committee believes that the introduction of an arbitrary scale for reporting scores is an unnecessary complication, and that reporting them on the same scale on which ratings are made will make them more immediately interpretable and meaningful. However, a minority of the Committee, and several consultants, believe that a separate metric is preferable to avoid confusion for applicants and others who may not be highly familiar with the rating and reporting systems.

Recommendation 9: Scores should be reported on the eight-point scale used by reviewers in making the original ratings. Scores should be reported with an implied precision commensurate with the information contained in the scores. Two significant digits are recommended.

5.4. A Global Rating of Scientific Merit

The above sections on rating and computing scores assumed that three scores, each based on ratings on scales of 0 to 7, would be computed for each application, representing the application's merit with respect to the three criteria -- ***Significance***, ***Approach***, and ***Feasibility***. However, the current system yields only a single score for each application, that of overall scientific merit.

The Committee believes that scores for each criterion capture extremely important information about applications, information that should be of significance in making funding decisions and in communicating feedback from reviewers to applicants. Therefore, the Committee recommends that these by-criterion ratings be made by reviewers and that by-criterion scores be the principal indices of scientific merit. The Committee held extensive discussions about whether a global score of scientific merit should be reported in addition to by-criterion scores. The Committee found attractive the concept of simply delivering a profile of three scores to ICs and the applicant. In such a case, if an IC desired a global measure of merit, it would be free to compute one from the three component scores using any algorithm that it deemed appropriate. Alternatively, the Committee understands that if global measures are to be widely computed and utilized, there are strong advantages in having such an index be computed in a uniform and consistent way throughout the NIH.

After extensive discussion with its outside board of experts, the Committee considered three algorithms for combining the component scores, two of them linear and one non-linear:

1) $O = (S + A + F) / 3$

2) $O = a*S + b*A + c*F$

3) $O = \text{Third Root of } S*A*F$

Where O is Overall scientific merit, S is the score on Significance, A is the score on Approach, and F is the score on Feasibility. Here the letters a, b and c represent weights for the respective criterion scores. The first formula is a simple arithmetic mean of the three scores where the criteria are equally weighted. The second formula is a sum of the criteria where different weights can be applied to the criteria representing different degrees of importance of the three criteria. The final formula represents a non-linear combination of the three criteria that is derived by taking the third root of the product of the three scores. In discussions of scoring algorithms, the view was often expressed that all three

criteria represented vital aspects of overall merit, and that a system should be developed in which a weakness on one criterion could not be compensated for by a strength on another. Linear formulae are generally compensatory in nature whereas non-linear formulae can be designed to be non-compensatory.

The Committee familiarized itself with literature on various linear models (Dawes, 1979), and a committee member, Dr. McGarvey, developed a Monte Carlo simulation (i.e., one based on values chosen to approximate reality as nearly as possible) of rating and scoring in which the various algorithms could be compared. His results are reported briefly in Section 5.5. They indicate that the three formulae given above correlate approximately equally with the "true values" of scientific merit as assigned to the simulated applications by the Monte Carlo system. Given that result, the Committee opts for simplicity and familiarity and recommends that, if a global score is to be computed, that it be the simple, unweighted arithmetic mean of the three criterion scores. Nevertheless, the simulations done to date show that the multiplicative model may have some small predictive advantage over the linear ones. Thus, the Committee recommends that additional studies be done to explore various formulae using a variety of initial conditions.

Recommendation 10: If a single score is required that represents overall merit, it should be computed from the three criterion scores using an algorithm that is common to all applications. The Committee favors the arithmetic average of the three scores; however, an appropriately constituted group should test and choose the algorithm to be used.

5.5. Simulations

In order to test the recommendations for standardizing scores and combining criterion scores into a global score, a Monte Carlo simulation was created that would allow the committee to vary a number of the characteristics of the system and observe their results in the final distribution of scores.

The system simulates a study section with various numbers of reviewers and various numbers of applications. "True" values of each of the three rating criteria are arbitrarily defined for each application. The distributions of ratings on each criterion in the applications "reviewed" can be varied as can the degree of agreement among the reviewers on the applications. The simulation assumes that the reviewers make their ratings on an seven-point scale and that the ratings are standardized within rater for each criterion. The simulation distinguishes between the "true" value of each application on each criterion and the ratings given by the raters so that the accuracy of the study section in identifying the best applications can be observed as a function of various initial conditions. Finally, the system can combine the criterion ratings into a global rating of scientific merit in any

number of different ways, thereby evaluating the ability of various algorithms to identify the best applications. The system is available for use by anyone interested in particular sets of initial conditions and algorithms for combining scores. (A more extensive description of the simulation system and some of the simulations performed are available in a separate document. Please contact Dr. William McGarvey for a copy at bm50b@nih.gov.)

The simulations performed to date have focused on the accuracy with which scores derived from combining ratings in various ways can discriminate superior applications from non-superior applications. They also have evaluated the simple product-moment correlations between these scores and the pre-determined values of scientific merit for these applications. Three algorithms for combining criterion scores into global scores were used (see section 5.4). For the weighted average, Significance was given a weight of 3.0, Approach was given a weight of 2.0, and Feasibility was given a weight of 1.0. When 10,000 scores were generated using each algorithm and correlated with the corresponding true values, the unweighted average correlated 0.68 with the predetermined standard, the weighted average correlated 0.69, and the product of the three scores correlated 0.78 with the true values. It should be noted that the absolute magnitudes of the correlations are a function of the initial parameters built into the simulation and therefore are somewhat arbitrary; however, their relative magnitudes are important and would indicate that the multiplicative algorithm is superior to the other two when correlation is taken as the measure of accuracy.

One approach to assessing accuracy of prediction is to partition the set of "true" values into those that would be prime candidates for funding and those that would not, and then to calculate the numbers of "hits" attributable to each algorithm. An advantage of this approach is that it also allows calculation of the relative numbers of each type of erroneous identification (false positives and false negatives). The simulation was run with 20 reviewers and with five reviewers to evaluate the effect that study section size has on the accuracy of scores obtained.

A prime candidate for funding was arbitrarily defined as an application that had a true value of at least 5.0 on each of the three criteria. Alternatively, an application was considered as having been identified empirically as a prime candidate if the computed global score fell approximately in the top 20 percent of the distribution of scores.

The results for the six simulations (three algorithms for scoring times two study section sizes) are as follows:

- For 20 reviewers the total number of hits (true positives plus true negatives) varied from 94.65 percent for the weighted average algorithm to 94.05 percent for the simple average algorithm.

- For five reviewers, the hit rate dropped to values ranging from 92.65 percent for the multiplicative algorithm to 92.34 percent for the simple average.

- False negatives and false positives were roughly balanced, with false positives being somewhat more numerous.

- The hit rates for 5 reviewers generally ran two to three percentage points less than the hit rates for 20 reviewers.

The immediate conclusion to be drawn from this simulation is that the accuracy of the system proposed is relatively robust across variations in the algorithms used to combine the criterion scores into an overall score, as well as in variations in numbers of raters. The former result is consistent with the findings of Dawes, Faust and Meehl (1989) and Dawes (1979) who found that within the class of linear models, the exact algorithm used did not have a large effect on its predictive value.

6. Summary of Proposed Procedures for Program Staff

Currently, program decision-making on grant awards varies across ICs. At some ICs, each individual division, branch or program cluster has its own budget and makes funding recommendations to the IC director and council that are generally sustained. At other ICs, one overall funding list is developed and the IC director determines where to draw the funding line. At still other ICs, the process is similar, except that a portion of the RPG funds are committed to strict percentile-based funding while another, generally much smaller portion is reserved for grants to be funded on the basis of high program relevance or portfolio balance. In all cases, the ICs rely on the recommendations provided by the reviewers, but in many cases the recommendations have been translated into a strict funding plan based on absolute percentiles or priority scores (for RFAs). Of major concern is not only the appearance, but also the reality, that ICs are largely - deferring their funding decisions to scientific review groups.

The approach under consideration by the RGA Committee not only addresses novel approaches to providing scientifically solid and meaningful review information to the ICs, it also provides for a system where programmatic relevance and IC mission can be considered in a more direct way within the framework of investigator-initiated research. Providing disaggregated scores and simplifying the numerous, and artificial, levels of scientific "bins" that both percentiled and priority score-based applications are currently assigned would result in more useful and programmatically flexible information for ICs. This system will allow ICs to better carry out their missions through the development of funding plans that take into account not only meaningful scientific merit, but also meaningful priorities.

The degree to which the new scoring system would alter the ways that program decisions are made would vary by ICs. In those ICs where each individual division, branch or program cluster has its own budget and makes funding recommendations to the IC Director and Council, changes would be less dramatic, since that funding decisions in those ICs are already based upon consideration of all factors -- priority scores, summary statements, program needs, and portfolio balance. Nevertheless, funding decisions at those ICs would still be based upon more useful information in the form of scores for the three disaggregated criteria.

In ICs where funding decisions rely primarily on percentiles or priority scores, alternative strategies for decision-making will be needed because it is anticipated that more applications in the fundable range will be identified as having equivalent scientific merit in the new scoring system. Although this is not anticipated to cause concern within the ICs, decision-making guidelines will need to be developed.

7. Implementation

The recommendations in this report cover a very broad scope of the initial review process, and any implementation of them must be accomplished in a way that will be minimally disruptive to the peer review system as it currently exists. In fact, implementation of some of them can proceed independently of the others. Ultimately, some may be implemented and others may not. Therefore, the Committee wishes to comment on the way in which it sees the recommendations relating to each other.

1) The revised review criteria could be implemented without implementing any other recommendation. If they are implemented, they could be used with or without implementing the procedures for using them and for rating by criterion. Critiques could be written by criterion without rating by criterion.

2) The rating scale could be changed without changing anything else.

3) The analysis and recommendations in Section 5 (Summarizing and Reporting Ratings) could be applied to any scale measuring any aspect of scientific merit. The last two recommendations apply only if applications are rated by criterion.

The Committee recommends that the revised review criteria be implemented as soon as possible. They could be used with the current review procedure, or they could be used with a by-criterion procedure and with or without a separate score being given to each criterion. The Committee believes that, at a minimum, the revised criteria could simply be substituted for the current criteria with little retraining of reviewers necessary.

All other recommendations will require additional study and testing before they are introduced into the regular initial review process. Although they may take varying amounts of time to be developed to the point of routine implementation, the Committee suggests that it may minimize the difficulty of transition to institute all changes (other than the review criteria) at a single time.

8. Evaluation

An important process in the implementation of a large-scale activity, such as the annual review of 30,000 grant applications, is to monitor and evaluate the effectiveness of the activity. Such an evaluation program is within the intent of the 1% program evaluation setaside funds within the Public Health Service. The Committee recommends that sufficient evaluation funds be provided to the Office of Extramural Programs to award contract(s) for the regular evaluation of the effectiveness of the process of review of grant applications. Evaluation contracts would involve studies of how grants are rated by setting up "shadow" study sections to review applications using alternative rating techniques. For example, raters in the shadow study section might be asked to give both a global independent rating in addition to the disaggregated criteria and compare the rating of the identical applications with those of the DRG study section.

9. In Closing...

This report has discussed a number of recommendations for changing the procedures by which initial review groups review, critique, and rate applications. Each of the recommendations flows from a basis in the literature of psychological and/or statistical measurement. Although the Committee made some effort to pilot test or simulate the systems it is recommending, every suggested change should be subjected to additional scrutiny and comment by all stakeholders in the review process and many should be pilot tested. (For example, the shift from percentiling to standardizing will involve further definition of the reference population in all possible situations.) The Committee believes that the details of implementation must be left to subsequent groups after this report has been circulated and subjected to comment, and after appropriate pilot testing and evaluation have been completed.

10. Solicitation of Comments

Any or all of the recommendations in this report could conceivably be implemented as part of the peer review process. We are currently considering the pros and cons of each recommendation, and the positive and negative impacts that each could have on the peer review system and on other aspects of the awarding of research grants at NIH. Comments may be sent to DDER@NIH.GOV until October 1, 1996.

11. References

Cicchetti, D.V., Showalter, D., and Tyrer, P.J. The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement,* 1985, *9*, 1, 31-36.

Dawes, R.M. The robust beauty of improper linear models in decision making. *American Psychologist,* 1979, *34*, 7, 571-582.

Dawes, R.M., Faust, D., and Meehl, P.E., Clinical versus actuarial judgment. *Science*, 1989, *243*, 1668-1674.

Knaus, W.A., Wagner, D.P., and Lynn, J., Short-term mortality predictions for critically ill hospitalized adults: Science and ethics. *Science*, 1991, *254*, 389-394.

Landy, F.J. and Farr, J.L., Performance Rating. *Psychological Bulletin*, 1980, *87*, 1, 72-107.

Miller, G. A., The magic number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 1956, *63*, 81-97.

Zuckerman, M. General and situation-specific traits and states: New approaches to assessment and other constructs. In M. Zuckerman and C. D. Spielberger (Eds.), *Emotions and anxiety: New methods and applications.* Hillsdale, N.J: Lawrence Erlbaum Associates. 1976, 133-174.

12. Committee Roster

Arkes, Hal; NSF (Now at Ohio University)
Atwell, Constance; NINDS
Bernick, Niles; OER
Coates, Paul; NIDDK
Heilman, Carole; NIAID
Jobe, Jared; NIA
Jordan, Elke; NCHGR
Levitin, Teri; NIDA
McGarvey, William; DRG (Now at OER)
Rawlings, Sam; DRG
Stamper, Hugh; NIMH (Co-Chair)
Stolz, Walter; NIDDK (Co-Chair)
Tingley, Dianne; NIAID

13. Consultants

May 17, 1996 (Revised)

Appelbaum, Mark; Vanderbilt University
Bakeman, Roger; Georgia State University
Dawes, Robyn; Carnegie-Mellon University
Goldberg, Lewis; Oregon Research Institute
Nesselroade, John; University of Virginia
Sechrest, Lee; University of Arizona