Statistical analysis of the National Institutes of Health peer review system

Valen E. Johnson*

University of Texas M.D. Anderson Cancer Center, 1400 Pressler Street, Unit #1411, Houston, TX 77030

Communicated by James O. Berger, Duke University, Durham, NC, May 15, 2008 (received for review February 17, 2008)

A statistical model is proposed for the analysis of peer-review ratings of R01 grant applications submitted to the National Institutes of Health. Innovations of this model include parameters that reflect differences in reviewer scoring patterns, a mechanism to account for the transfer of information from an application's preliminary ratings and group discussion to final ratings provided by all panel members and posterior estimates of the uncertainty associated with proposal ratings. Application of this model to recent R01 rating data suggests that statistical adjustments to panel rating data would lead to a 25% change in the pool of funded proposals. Viewed more broadly, the methodology proposed in this article provides a general framework for the analysis of data collected interactively from expert panels through the use of the Delphi method and related procedures.

hierarchical model | item response model | latent variable model | ordinal data

Every year, the National Institutes of Health (NIH) spend more than $22 billion to fund scientific research (1). Approximately 70% of these funds are awarded through a peer-review process overseen by the NIH Center for Scientific Review (CSR). Despite the vast sum of money involved, the absence of statistical methodology appropriate for the analyses of peer-review scores generated by this system has precluded the type of detailed assessment applied to other national health and educational systems (2, 3). As a consequence, statistical adjustments to account for uncertainties and biases inherent to these scores are not made before funding decisions. To address this deficiency, this article examines the properties of these ratings and proposes methodology to more efficiently extract the information contained in them.

It is useful to begin with a brief review of the NIH peer-review system. Upon submission to the NIH, most grant applications (e.g., R01, R03, R21, etc.) are assigned to a study section within an Integrated Review Group (IRG) for review, and to an NIH Institute and Center (IC) for eventual funding. IRG study sections typically contain ≈30 members and review ≈50 grant applications (proposals) during each of three annual meetings. Because it is impractical for every member of a study section to review every application, between two and five reviewers are typically assigned to read and score each application before the study section convenes. In the sequel, these individuals are called the proposal's "readers," and the scores they assign before a study section convenes are called "pre-scores." Proposals are scored on a 1.0–5.0 scale in increments of 0.1 units, with 1.0 representing the best score. When the study section convenes, the scientific review officer (SRO) and the study section chair suggest a list of proposals that might be "streamlined." Based on their pre-scores, proposals on this list are viewed as unlikely to receive fundable priority scores and, if no one in the study section objects, are not considered further. The remaining proposals are discussed and scored by all members of the study section.

Readers of a grant application begin the discussion by announcing their pre-scores and summarizing the proposal for other members of the study section, most of whom will not have read it. After these summaries, there is an open discussion of the application. Proposal readers then state their "post-scores" for the application, and all other members of the study section (i.e., the proposal's nonreaders) also score the proposal. Nonreaders are required to either score the proposal within 0.5 units of the range of scores established by reader post-scores or provide a written statement to the SRO explaining why they scored the proposal outside of that range. Scores received from all study section members are then averaged to obtain the proposal's priority score. In "established" study sections, priority scores are converted to a percentile ranking through a comparison with recent priority scores from other grant applications scored within that study section. In newer study sections or special emphasis panels (i.e., panels that are convened to rate a limited number of proposals), percentile scores are calculated by comparing the proposal's priority score to established norms. Finally, proposal percentile ratings are used by ICs to determine which applications will be funded. Although the exact criteria by which ICs use these percentiles to make funding decisions vary by the IC, funding decisions are thought to be highly correlated with percentile scores.

In this article, I propose statistical methodology to account for the effect of the selection of readers on a proposal's final percentile score, quantify the uncertainty associated with the percentile scores, and demonstrate how such uncertainties can be incorporated into a decision-theoretic framework to improve the probability that the greatest proportion of top proposals are funded. Viewed more generally, methods developed in this article extend existing statistical methodology for the analysis of multirater ordinal data (4–7) and item response data (8–12) to provide a framework for the analysis of panel rating data collected by using the Delphi method and related interactive rating schemes (13).

The data that form the basis for this study were collected as part of a contract awarded to the author by the CSR in November 2004. As part of that study, all preliminary and final reader scores and nonreader scores for all R01 grant proposals submitted to the NIH and reviewed under the auspices of the CSR over two review cycles (June and October 2005) were collected and redacted.

Description of Data

Ratings for 18,959 R01 proposals rated by 14,041 reviewers in 744 study sections (including special emphasis panels) were available for analysis. Fig. 1 displays a histogram of all scores, including reader pre-scores and post-scores, and nonreader

Author contributions: V.E.J. designed research, performed research, analyzed data, and wrote the paper.

The author declares no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: Dr. Johnson will provide the data in ASCII format upon request.

*E-mail: vejohnson@mdanderson.org.

This article contains supporting information online at www.pnas.org/cgi/content/full/0804538105/DCSupplemental.

© 2008 by The National Academy of Sciences of the USA

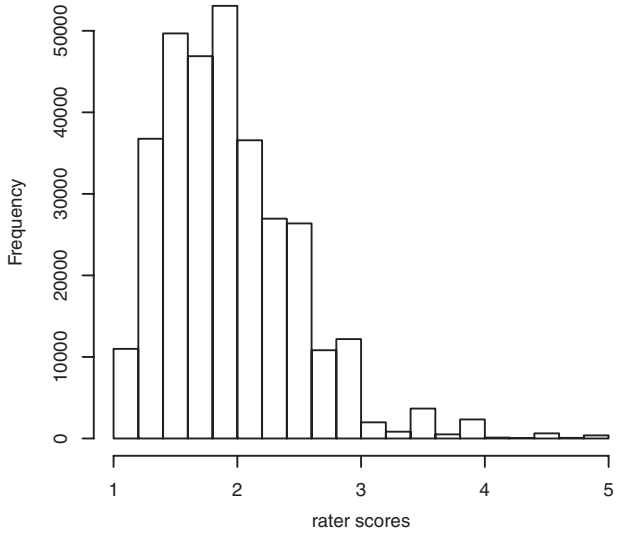

Fig. 1. Histogram of rater scores (including reader pre-scores and post-scores, and nonreader scores) assigned to R01 proposals.

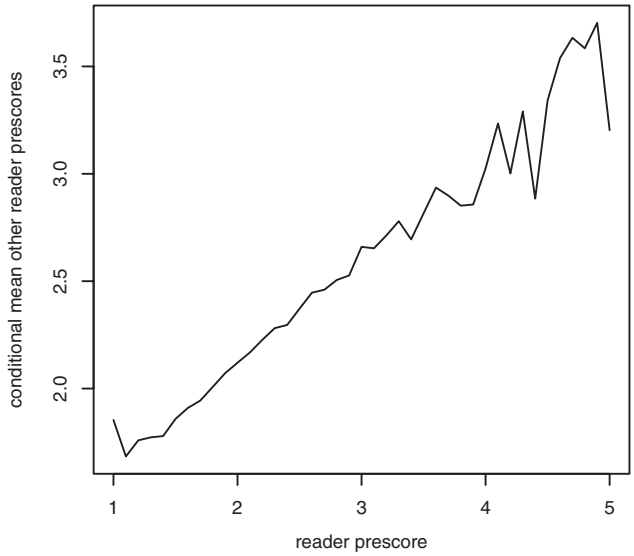

Fig. 2. Plot of the conditional mean pre-scores assigned by other readers versus single reader pre-scores.

scores. Table 1 provides a summary of the mean and standard deviation of the rater scores.

Several interesting features of the data are apparent from Fig. 1. Among these is a tendency for reviewers to use two distinct scales to score proposals. The first scale, nominally assumed by the CSR, runs from 1.0 to 5.0 in increments of 0.1 units. The second scale, used more frequently for less competitive proposals, runs from 1.0 to 5.0 in increments of 0.5 units. Evidence for the operation of these dual scales is provided in Fig. 2, in which the conditional means of reader pre-scores are displayed as a function of the prescore assigned to a proposal by a single reviewer. The relation between a reader prescore and the mean of other reader pre-scores for the same proposal is nearly linear between ≈1.1 and 3.0, but, outside of that range, the relationship is not monotonic. For example, among proposals that receive one prescore of 5.0, the mean of the remaining pre-scores is 3.2; for proposals receiving a prescore of 4.9, the mean of the remaining pre-scores is 3.7. Although not a central focus of this article, these observations suggest that a 20-point scale, anchored at an "average" rating of 10, might be better supported by current rating procedures. Such a scale would nominally provide a 10-point scale for nonstreamlined proposals.

Results

I used a latent variable model (14, 15) to formally describe the relation among application merit, reader pre- and post-scores, and nonreader scores. Within this model, reader pre-scores were assumed to represent independent assessments of application merit, whereas reader post-scores and nonreader scores were assumed to represent weighted averages of information elicited during the proposal discussion and the scores of (other) proposal readers. I used a continuous-valued latent variable μ_i to represent the merit of the ith application. The resulting model was then used to estimate the effects of reader biases and to assess the uncertainty in final proposal rankings. A description of this statistical model is provided in the supporting information (SI).

Table 1. Summary statistics for R01 proposal rater scores

	pre-scores (all)	pre-scores (not streamlined)	post-scores	non-reader scores
sample mean	2.21	1.88	1.90	1.96
std. deviation	0.77	0.51	0.49	0.50

Columns provide the mean and standard deviations of reader pre-scores for all proposals, reader pre-scores for proposals that were not streamlined, reader post-scores, and nonreader scores.

Adjustments for Reader Bias. Demonstrating the benefit of corrections for reviewer bias is difficult because true proposal merits are not known. For this reason, I examined the effectiveness of bias corrections in two stages. First, I performed a cross-validation study that used only reader pre-scores. Because reader pre-scores can be considered to be conditionally independent, they can be analyzed without modeling the complex structure among their values, reader post-scores, and nonreader scores. Therefore, a comparison of the model-based prediction errors based on reader pre-scores to the NIH prediction error provides an indication of the effectiveness of corrections for reader biases and a partial model validation. Second, I applied the full statistical model to all rater scores to illustrate the impact of reader bias on the final estimates of the proposals' merits.

I implemented the cross-validation experiment by first splitting reader pre-scores into two samples, randomly assigning 90% of the scores to a training sample and assigning the remaining 10% to a test sample. I used the training data to estimate model parameters. The posterior means of merit parameters for the proposals were then converted back to the original rating scale and were used to predict pre-scores in the test sample. The mean squared error for these predictions was 0.373.

In the NIH scoring system, proposal merit is estimated by the sample mean of the raters' scores. Thus, the estimate of a proposal's merit based on the training sample is the sample mean of the training sample pre-scores. The mean squared error of the corresponding prediction of pre-scores in the test sample was 0.413. Use of the statistical model to predict reader pre-scores in the test sample thus reduced the mean squared error of prediction by ≈10%.

The improvement in mean squared error enjoyed by the model-based estimate can be attributed primarily to the estimation of parameters that represent rater biases, or the tendency of some raters to score proposals more stringently than others.

When propagated through the full statistical model for reader post-scores and nonreader scores, these effects can be quite dramatic. For example, consider the posterior estimates of the proposal merits listed in Table 2. These proposals represent the

PSYCHOLOGY

STATISTICS

Table 2. Ratings of top fifteen proposals from the first study section in the NIH data set

Rank	$\bar{\mu}$	$\bar{y}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	$
1	-2.14	1.18	-	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	350
2	-1.54	1.25	1.0	-	.01	.00	.00	.00	.00	.08	.00	.00	.00	.00	.01	.00	.00	300
3	-1.26	1.35	1.0	.99	-	.41	.08	.05	.00	.17	.00	.00	.00	.00	.01	.00	.00	400
4	-1.24	1.3	1.0	1.0	.59	-	.14	.10	.01	.18	.00	.00	.00	.00	.01	.00	.00	200
5	-1.12	1.34	1.0	1.0	.92	.86	-	.45	.07	.20	.00	.00	.00	.00	.01	.00	.00	250
6	-1.11	1.34	1.0	1.0	.95	.90	.55	-	.08	.20	.00	.00	.00	.01	.02	.00	.00	450
7	-0.98	1.35	1.0	1.0	1.0	.99	.93	.92	-	.24	.00	.03	.00	.04	.03	.00	.00	450
8	-0.86	1.61	1.0	.92	.83	.82	.80	.80	.76	-	.40	.45	.30	.38	.22	.03	.02	400
9	-0.7	1.53	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.60	-	.65	.33	.52	.28	.00	.00	350
10	-0.69	1.65	1.0	1.0	1.0	1.0	1.0	1.0	.97	.55	.35	-	.29	.42	.27	.18	.18	350
11	-0.66	1.59	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.70	.67	.71	.00	.60	.33	.01	.00	300
12	-0.65	1.56	1.0	1.0	1.0	1.0	1.0	.99	.96	.62	.48	.58	.40	-	.31	.21	.20	400
13	-0.46	1.68	1.0	.99	.99	.99	.99	.98	.97	.78	.72	.73	.67	.69	-	.51	0.50	350
14	-0.45	1.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.97	1.0	.82	.99	.79	.49	-	0.46	150
15	-0.45	1.82	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.98	1.0	.82	1.0	.80	.50	.54	-	200

Posterior means of proposal merit parameters are listed in column two of the table ($\bar{\mu}$), while the mean priority score rating is provided in column three ($\bar{y}$). Columns 4–18 provide the posterior probability that one proposal is better than each of the others. Hypothetical proposal costs (in thousands of dollars) are listed in the final column.

top 15 applications selected from a study section that reviewed 99 proposals over the two cycles for which data were collected. Proposal rankings were based on the posterior means of the μ_j, which are listed in column two of the table ($\bar{\mu}$). The sample mean of reader post-scores and nonreader scores are listed in column three ($\bar{y}$). Columns 4 through 18 provide the posterior probabilities that each proposal had higher merit than each of the other proposals. Note that there is substantial disagreement between the ordering of proposals obtained from the statistical model and the raw priority score averages. Because these differences are so great, it is helpful to examine their source. To this end, consider the most extreme example from Table 2, the eighth proposal.

This proposal had a posterior mean estimate of $\bar{\mu}_8 = -0.86$, which was based on four reader pre-scores of 1.2, 1.8, 2.5, and 1.4. The third reader, who assigned this proposal a prescore of 2.5, assigned pre-scores of 2.6, 2.8, and 2.2 to the three other proposals he read, and as a consequence was estimated to have a relatively large, positive bias. Similarly, the second reader, who assigned the prescore of 1.8, also graded more stringently than average, assigning an average prescore of 2.0 to the 11 proposals she reviewed. The pre-scores assigned to this proposal by the first and fourth readers are even more unusual and were the lowest pre-scores that these readers assigned to any proposal. These two panel members prescored 7 and 12 grants, respectively, and assigned average pre-scores of 2.74 and 2.33. The reader post-scores of this grant application, in order from the first to the fourth reader, were an abstention, 1.8, 2.5, and 1.2. There was thus considerable disagreement among the readers concerning the merit of this proposal.

This discord carried over to the nonreaders of the proposal, who were split in their opinions. Ten of 22 nonreaders scored the proposal 1.2 or 1.3, whereas 7 of 22 scored the proposal 1.8 or higher.

The scores of this proposal thus reflect one obvious but important feature of the NIH scoring system: The scoring patterns of readers assigned to an application have a major impact on its final priority score.

Restricting attention only to the effects of rater biases, the model-based correction for these effects changed the rank of the eighth proposal from 13 to 8, or from being near the current NIH funding line to being under it. Applying similar corrections to proposals in all study sections suggests that corrections for rater biases would lead to a change in ≈25% of funding decisions. At a 15% funding line, 20% of funded proposals would be replaced by unfunded proposals if an account was made for the differences in reader scoring patterns. At a 10% funding line, this difference becomes ≈27%. In dollars, this translates to the redirected allocation of approximately $5 billion of grant funding every year.

Uncertainty in Proposal Ratings. Uncertainties associated with proposal orderings should also be considered when allocating research funds, particularly when uncertainty is great (2, 3, 16). To examine the importance of this factor, consider again the eighth-ranked proposal from Table 2. Because of the disparity of scores assigned to this application, it is difficult to accurately determine its relative merit. The posterior probability that it was better than the ninth-ranked proposal was estimated to be only 0.60, and the posterior probabilities that it was better than the 10th- and 11th-ranked proposals were 0.55 and 0.70, respectively. Yet there was a 24% chance that it was better than the seventh-ranked proposal and a 20% chance that it was better than the fifth- and sixth-ranked proposals.

These probabilities reflect another feature of the model-based estimates of each proposal's merit that is not captured by the sample mean of the priority scores. The actual merit of this proposal is not clear from the reader scores nor the nonreader scores; it could rank in the top four or five proposals from this study section, or it might only be among the top 10 or 15.

More generally, a statistical model to determine the merit of each proposal provides a mechanism for balancing the estimates of posterior uncertainty regarding the relative merit of proposals against the requested costs of proposals to arrive at more rational funding decisions. To understand how this might be accomplished, consider again the proposals summarized in Table 2.

Because the costs requested in the proposals in Table 2 are not available, hypothetical costs have been inserted into the final column of the table. For convenience, a proposal's total costs were assumed to be distributed between $200,000 and $450,000, based on the assumption that the average funding of an R01 proposal is approximately $350,000 (1).

In the absence of a formal utility function for proposal merit, let us assume that the NIH wishes to maximize the probability that the top, e.g., 13% of grant applications are funded under a fixed constraint on the available funding. Suppose further that $13 \times 350{,}000 = \$4.55$ million is available to fund a subset of the proposals listed in Table 2 and recall that 99 proposals were rated by this study section.

Without accounting for the uncertainty in proposal rankings, a natural funding decision would be to simply fund the top 13 proposals in the table. The combined cost of these proposals is 4.55 million dollars, and so this selection might appear to maximize the probability that the top 13% of proposals would be

funded. However, this choice does not account for the uncertainty associated with the estimates of the relative merit of proposals 13–15.

To account for the uncertainty in the relative merits of proposals, the numerical algorithm used to sample from the posterior distribution was also used to rank proposals for each sample generated from the posterior distribution. Based on these samples, it was possible to calculate the probability that each fundable subset of proposals (i.e., a group of proposals costing less than $4.55 million) contained the 13 top proposals. The posterior probability that proposals 1–13 were the 13 best proposals was thus calculated to be 17%.

Given the imposed cost constraints and noting that proposals 14 and 15 have the same total cost as proposal 13, an alternative funding decision would be to fund proposals 1–12, 14, and 15. The combined cost of these proposals is also $4.55 million. Perhaps surprisingly, the posterior probability that this set of proposals contains the 13 best proposals is 21%—nearly 24% greater than the probability achieved by the selection of proposals 1–13. Clearly, this selection of proposals would significantly increase the NIH's probability of funding the top 13% of proposals from within this study section.

This general approach for combining uncertainty and costs extends easily to different target levels of funding, or to funding decisions made for proposals pooled from several study sections. In addition to maximizing the probability that the top proposals are funded, using such an approach to balance costs against uncertainties would also have an additional benefit: It would decrease the costs requested in grant applications. In the current highly competitive funding environment, applicants would submit reduced budgets if they knew this would improve their chance of being funded.

Discussion

The statistical model proposed in this article illustrates the potential that exists for modeling rating data collected interactively from panels of experts. It accounts for differences in reviewer scoring criteria, provides a model for the sequential rating of items by various subsets of reviewers, and quantifies uncertainty associated with final proposal ratings. Numerous refinements to this model framework are clearly possible. For example, the model could be extended to account for differences between the weights assigned to ratings by primary reviewers, secondary reviewers, and discussants or for differences that might be explained by reviewer attributes [e.g., academic rank, gender, ethnicity, scientific review group (SRG) experience]. Indeed, entirely different classes of statistical models might alternatively be considered, and it would be worthwhile to assess the sensitivity of funding decisions to the particular model adopted. Within the context of NIH peer-review rankings, an issue that urgently requires additional study involves the impact of review group discussion on the final rankings of proposals. The approach taken here represents an extremely optimistic view of this "discussion effect." That is, systematic shifts in reader pre-scores to reader post-scores and nonreader scores were assumed to result from an implicitly unbiased glimpse of the true merit of a proposal manifested through group discussion. In practice, group dynamics and reviewer attributes probably play as important a role in such discussions as do the proposals' merits. Unfortunately, the data do not contain unambiguous information regarding the true value of review group discussions or the possible biases associated with them. Such information might be obtained, however, through an experimental study of the rating process itself.

Perhaps the simplest experiment that could be conducted to assess the validity of the discussion effect would be to set aside from SRG discussion a random sample of reader prescore and postscore information. Nonreader scores could subsequently be contrasted to omitted pre-scores, which (under the assumptions of the model above) could be corrected to provide unbiased estimates of the proposals' merits. The relative distribution of deviations of nonreader scores from omitted pre-scores and reported pre-scores would provide an indication of the extent to which a discussion of a proposal represents an independent assessment of its merit. Such analyses could be strengthened by examining the impact of individual reader attributes (e.g., academic rank, gender, years of SRG experience) on observed shifts of nonreader scores toward reported reader pre-scores. Ultimately, data collected from such experiments might be used to assess the tradeoff between the cost of conducting SRG meetings and the cost of collecting additional, independent ratings of applications.

There is, however, no ambiguity regarding the need for more sophisticated statistical analyses of NIH peer-review data. As the example in the previous section illustrates, variability inherent to rater scores, and differences in the criteria used by individual raters to assign scores to proposals, have an enormous impact on funding decisions. The statistical model proposed in this article—or a modification of it—should be applied by the NIH to account for these effects.

The primary technical difficulty associated with the implementation of this model stems from the estimation of model hyperparameters that are common to all SRGs. For example, estimation of hyperparameters that model the correlation of category thresholds across review groups requires the evaluation of the complete likelihood function, which depends on rating data collected from all study sections. However, collecting these data from all SRGs to estimate global model hyperparameters would delay the processing of summary scores. A practical implementation of the model would thus require both an off-line procedure to estimate and update the posterior distributions of global model hyperparameters based on past review cycle data, and the updating of the values (or summary statistics describing the posterior distribution of values) of a static set of global hyperparameters used concurrently in end-user software (17).

Implementation of such a system would likely change the pool of funded proposals by 25%; accounting for both requested costs and uncertainty in the relative merits of proposals would likely result in more than a 35% change. Explicitly accounting for cost in funding decisions would also result in a net decrease in the cost of the average proposal, which in turn would allow the NIH to fund more grant applications.

Methods

I used a Bayesian hierarchical statistical model to describe the process by which raters scored grant applications. Stages in the model hierarchy were specified sequentially according to the order in which scores for proposals were generated.

First-Stage Model. I modeled reader pre-scores using ordinal probit models (6, 18) defined by using latent variables. Letting μ_i denote the "true" merit of proposal i on an underlying measurement scale, r_j denote a "bias" term associated with the pre-scores assigned by reader j, γ_m denote a vector of category thresholds associated with IRG study section m, and $x_{i,j}^{\mathrm{pre}}$ denote the unobserved latent variable upon which reader j assigns prescore $y_{i,j}^{\mathrm{pre}}$ to proposal i, such a model was specified by assuming that

$$\begin{aligned} x_{i,j}^{\mathrm{pre}} &= \mu_i + r_j + \varepsilon_{i,j}^{\mathrm{pre}}, \\ y_{i,j}^{\mathrm{pre}} &= s \Leftrightarrow \gamma_{m,s-1} \leq x_{i,j}^{\mathrm{pre}} < \gamma_{m,s}. \end{aligned} \quad [1]$$

To establish the underlying scale of measurement, the μ_i values were assumed to be independently distributed as standard normal deviates.

A priori, biases attributable to raters and the error terms associated with the assignment of pre-scores to categories were assumed to be independent and distributed according to

$$r_j \sim N(\zeta, \tau^2) \quad \text{and} \quad \varepsilon_{ij}^{\mathrm{pre}} \sim N(0, \sigma_0^2).$$

PSYCHOLOGY

STATISTICS

The mean of rater biases ζ was included in the model to account for the fact that reader pre-scores have a lower mean value than either the rater post-scores or nonreader post-scores.

Second-Stage Model. In the next stage of the data generating process, I assumed that readers modified their pre-scores by using both the reported values of other reader pre-scores and the group's discussion of the proposal. The resulting reader post-scores were thereby represented as a weighted average of these three information sources.

The latent value $x_{i,j}^{\text{post}}$ assumed to be responsible for the generation of reader j's postscore of proposal i, y_{ij}^{post}, can be written as

$$x_{i,j}^{\text{post}} = u_{i,j}x_{i,j}^{\text{pre}} + v_i\mu_i + \sum_{k\varepsilon\mathcal{A}_i;k\neq i} w_{i,j,k}x_{i,j,k}^{\text{pre}} + \varepsilon_{ij}^{\text{post}}, \qquad [2]$$

where

$$y_{i,j}^{\text{post}} = s \Leftrightarrow \gamma_{m,s-1} \leq x_{i,j}^{\text{post}} < \gamma_{m,s}, \qquad [3]$$

and

$$u_{i,j} + v_i + \sum_{k\in\mathcal{A}_i;k\neq i} w_{i,j,k} = 1. \qquad [4]$$

The error terms $\varepsilon_{ij}^{\text{post}}$ were assumed to be independently distributed as $N(0,\sigma_1^2)$ random variables. Here, $\mathcal{A}_i$ denotes the set of reviewers who provided pre-scores for proposal i.

On the latent scale of measurement, the model specification described so far resembles a standard hierarchical model with a Gaussian error structure. Unfortunately, the usual Gaussian model does not provide an accurate representation of reader post-scores and nonreader scores at higher levels in the model hierarchy. This difficulty stems from the high proportion of reader post-scores that fall within the range defined by the reader pre-scores, and the even higher proportion of nonreader scores that fall within the range defined by the reader post-scores. There also is a tendency for nonreaders to assign scores that are identical to a reader postscore.

To account for these tendencies, the weights u_{ij}, v_i, and w_{ijk} were assumed to be generated from a Dirichlet model with a parameter vector containing a component a for each u_{ij}, a component b for each v_i, and a component c for each w_{ijk}. The distribution of hyperparameters estimated at higher levels in the model hierarchy make it likely that these weights are assigned values that are either close to 0 or 1; this permits the model to mimic the tendency of nonreaders to concentrate their scores around and between the scores recorded by the proposal's readers.

Another innovation of the statistical model involves the inclusion of the term $v_i\mu_i$ in the weighted average defining the latent variable $x_{i,j}^{\text{post}}$ (Eq. **2**). The purpose of this term is to model systematic shifts between reader pre-scores and reader post-scores that result from a proposal's discussion. In the construction of this term, v_i weights μ_i, the parameter that represents the true merit of the proposal. That is, the model implicitly assumes what might be regarded as the ideal situation from the NIH's standpoint. Alternative assumptions regarding the distributions of these weights can be incorporated into the model framework, but for the purposes of this article the NIH's ''ideal'' was assumed. It is important to note, however, that the rating data themselves cannot be used to validate this assumption in the absence of an external ''gold standard'' for relative proposal merits.

The values of the hyperparameters a, b, and c determine, respectively, the average relative weights that readers assign to their own pre-scores, the proposal discussion, and the pre-scores of other readers when determining their final postscore ratings.

Third-Stage Model. The model for nonreader scores $y_{i,j}^{\text{non}}$ is similar to the model specified for reader post-scores $y_{i,j}^{\text{post}}$, except that nonreader scores were assumed to be based on a latent variable $x_{i,j}^{\text{non}}$ that represents a weighted average of reader post-scores and proposal merit. That is, the model for nonreader scores was obtained by replacing Eq. **2** with

$$x_{i,j}^{\text{non}} = v_i\mu_i + \sum_{k\in B_i} w_{i,j,k}x_{i,j,k}^{\text{post}} + \varepsilon_{i,j}^{\text{non}}, \qquad [5]$$

and modifying Eqs. **3** and **4** accordingly. The weights appearing in Eq. **5** were defined similarly to those used to model reader post-scores.

Further description of higher-level model structures [including the prior distributions imposed on model hyperparameters (γ_m, a, b, c, σ_0^2, σ_1^2, σ_2^2, τ^2)], along with model diagnostics and a brief description of the numerical algorithm used to fit this model to the peer-review data, is provided in the SI.

ACKNOWLEDGMENTS. I thank James Berger and two referees for constructive comments and suggestions that significantly improved the manuscript.

1. Office of Budget, National Institutes of Health (2007) Summary of the FY 2008 President's Budget. Available at http://officeofbudget.od.nih.gov/PDF/Press%20info-2008.pdf.
2. Goldstein H, Spiegelhalter DJ (1996) League tables and their limitations: Statistical issues in comparisons of institutional performance. *J Roy Stat Soc* 159:385–443.
3. Bird SM, *et al.* (2005) Performance indicators: Good, bad, and ugly. *J Roy Stat Soc* 168:1–27.
4. Albert JA, Chib S (1993) Bayesian analysis of binary and polychotomous response data. *J Am Stat Assoc* 88:669–679.
5. Johnson VE (1996) On Bayesian analysis of multirater ordinal data: An application to automated essay grading. *J Am Stat Assoc* 91:42–51.
6. Johnson VE, Albert JA (1999) *Ordinal Data Modeling* (Springer, New York).
7. Ishwaran H (2000) Univariate and multirater ordinal cumulative link regression with covariate specific cutpoints. *Can J Stat* 28:715–730.
8. Verhelst N, Verstralen H (2001) IRT models for multiple raters. *Essays on Item Response Theory*, eds Boosma A, van Duijn M, Snijders T (Springer, New York), pp 89–108.
9. Wilson M, Hoskens M (2001) The rater bundle. *J Educ Behav Stat* 26:283–306.
10. Patz RJ, Junker BW, Johnson MS, Mariano LT (2002) The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *J Educ Behav Stat* 27:341–384.
11. Skrondal A, Rabe-Hesketh S (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models* (CRC, Boca-Raton, FL).
12. Mariano LT, Junker BW (2007) Covariates of the rating process in hierarchical models for multiple ratings of test items. *J Educ Behav Stat* 32:287–314.
13. Rowe G, Wright G (1999) The Delphi technique as a forecasting tool: Issues and analysis. *Int J Forecast* 15:353–375.
14. Bollen KA (2002) Latent variables in psychology and the social sciences. *Annu Rev Psychol* 53:605–634.
15. Borsboom D, Mellenbergh GJ, van Heerden J (2003) The theoretical status of latent variables. *Psychol Rev* 110:203–209.
16. National Research Council (2005) *Strengthening Peer Review in Federal Agencies That Support Education Research* (National Academies Press, Washington, DC).
17. Kolen, MJ, Brennan RL (1995), *Test Equating: Methods and Practices* (Springer, New York).
18. McCullagh P (1980). Regression models for ordinal data. *J Roy Stat Soc Ser B Method* 42:109–142.