New Decision Tool to Evaluate Award Selection Process

Richard Thornley, Matthew W. Spence, Mark Taylor, and Jacques Magnan
Alberta Heritage Foundation for Medical Research

Abstract

This article describes an Alberta Heritage Foundation for Medical Research (AHFMR) initiative to enhance the review process for its training awards using a new tool based on the ProGrid® decision-assist software. Changes that the AHFMR made are analyzed, as is the manner in which the review system was developed and refined, the initial assessment of the value added for applicants, reviewers, and the funding decision-maker, and the implications of this approach for the future. Implementation of the new tool resulted in a number of modifications to the AHFMR's review process in the areas of definition, rationality, fairness, timeliness, and responsiveness. The new process also provides the foundation with increased capacity for performance measurement and program evaluation.

Introduction

Established by the Government of Alberta in 1979, the Alberta Heritage Foundation for Medical Research (AHFMR) supports health research at Alberta universities and other research-related institutions. The foundation supports nearly 230 faculty-level researchers recruited from Alberta and around the world, and approximately 500 researchers-in-training (i.e., summer students, graduate students, and post-doctoral fellows, collectively known as trainees). The AHFMR's gross expenditure for fiscal year (FY) 2000-2001 was approximately $53 million, of which $6.7 million (12.6%) was committed to the funding of trainees.[1] This article describes the foundation's initiative to improve the peer review process for its competitive training awards.

Peer review is frequently used for both ex ante and ex post evaluation of the quality of the scientific enterprise (Geisler, 2000; Kostoff, 1992; Luukkonen-Gronow, 1987; United States General Accounting Office, 1997). Ex ante evaluation assesses quality in advance of performance, as in the case of applications for research funding. Conversely, ex post evaluation assesses quality retrospectively, as in the case of papers submitted to scientific journals. The case described here

Author's Note: The early drafts of this paper were prepared for internal presentations to various AHFMR advisory committees. Thanks to the reviewers of the SRA Journal for their comments and suggested edits.
Correspond to Mr. Thornley at the Alberta Heritage Foundation for Medical Research, Suite 1500, 10104-103 Avenue, Edmonton, Alberta T5J4A7, Canada Email: Richard.thornley@ahfmr.ab.ca

[1] *All amounts are in Canadian dollars.*

entails ex ante review of applications for funding, to anticipate the future performance of research trainees.

The AHFMR's original review process for training award applications considered three general criteria: (a) the quality of the candidate, (b) the appropriateness of the proposed research environment, and (c) the merit of the proposed research project. Applications were rated following a multiple-step committee process on a scale of 0 to 5, the single score representing an aggregation of performance in relation to all criteria. Zero is considered an unacceptable application whereas a score of 5 is an outstanding application. This approach was used by the foundation to review applications for its training awards until the end of FY2000, when the foundation piloted the new process described here.

Geisler (2000) suggested that peer review should be well-defined, rational, fair, timely, cost-effective, anonymous, and responsive. While most of these general characteristics were reflected in the AHFMR's original review process for its training awards, a number of specific issues provided the incentive for the foundation to try to improve the process.

First, the number of proposals submitted was increasing and there was a need to more efficiently evaluate them. In FY1997, the AHFMR received 182 applications for full-time studentships, as compared to 276 in FY2000 and 307 in FY2001. This resulted in the need for more reviewers, most of whom were reporting that they had increasingly less time to devote to such activities. Also, the increase in proposals meant that committees were faced with extending the duration of their meetings or spending less time reviewing each application, neither of which was considered to be a desirable alternative.

This issue was complicated by an increase in turnover on the foundation's review committees. In general, this may have been in response to reviewer fatigue, a recent and widespread phenomenon in the research funding sector resulting from a proliferation of requests to individuals to sit on review panels (Brzustowski, 2000a; Brzustowski, 2000b; Cunningham, Boden, Glynn, & Hills, 2001; Smith, 2001). There was a sense that turnover resulted in less consistency in the application of criteria within and between competitions, and an increased administrative burden in recruiting and training committee members.

Two trends relating to scores awarded to applications also influenced the AHFMR's decision to redesign its review process. In theory, the overall score awarded to each application represented an integration of all parts of the application; however, in practice each reviewer's interpretation resulted in variable weighting of different criteria. For example, one reviewer might value the quality of the candidate significantly more than that of the proposed research environment or the proposed research. Consequently, an application from a promising candidate working on a less promising research project might receive a higher score from this reviewer than it would from a reviewer who equally emphasized all categories of criteria. While this different weighting between individual reviewers was correctable within the scope of the overall meeting discussion, there were also indications that a committee, as a result of reviewing hundreds of proposals, would in some cases alter its approach within competitions (i.e., resulting in inconsistently applied review criteria between groups of applications). Also, as committee members became familiar with review criteria, the cut-off point between funded and rejected applications began to creep upwards;[2] the system began to result in more limited discrimination (approximately 75% of proposals were rated between 3 and 4.25). For these reasons, the AHFMR recognized that a process that supported a more consistent application of criteria between reviewers and within competitions would be beneficial.

Finally, candidates had requested more feedback on a routine basis from the AHFMR about the review of the applications. In

[2] This common phenomenon was observed in Hodgson's (1995) study of grant proposal reviews at the Heart and Stroke Foundation of Ontario and has also been reported by other agencies, such as the National Institutes of Health (NIAID Council, 1999.)

general, changes were sought that would increase the transparency of the review process and result in this outcome.

The Intervention

The primary intervention was the incorporation of the ProGrid® decision-assist software into the review process for competitions for training awards.[3] The foundation had prior experience with this tool and recognized that an extension of the methodology to the competitions for training awards had potential. Also, other Canadian research organizations, such as the Canadian Foundation for Innovation (CFI) and Natural Sciences and Engineering Research Council (NSERC), used this tool.[4]

The ProGrid approach requires the developer to articulate explicit criteria upon which decisions will be based. In the case of the AHFMR, a number of factors were identified that were thought to be predictive of the performance of trainees in the proposed research environment (see Table 1). These performance factors were attributed to one or both of two independent performance criteria: the characteristics of the candidate and the characteristics of the candidate's proposed research environment. It was felt that these two general criteria were those most indicative of the future performance of research trainees. Variations of these criteria are used by other R&D funding agencies in Canada.[5]

Constructing Tools

Some of the performance factors are clearly attributable to one of the two major independent criteria. For example, the academic record relates directly to the characteristics of the candidate and not to the candidate's proposed research environment. The supervisor's research record, conversely, relates primarily to the characteristics of the research environment and little to the characteristics of any given candidate, especially for those candidates who have yet to join the supervisor's research group. However, a number of linking factors, such as the role of the trainee, were identified as contributing to varying degrees to the two general criteria. The end result of this process was the Trainee Evaluation Matrix (see Table 1) which represents an articulation of the foundation's values, priorities, and expectations of the training award candidates and their supervisors. The factors identified in the matrix represent those that had been used in the reviewers' discussions in the original review process. In essence, the matrix articulated these factors more explicitly than the original process.

A *language ladder* was then constructed for each performance factor in the matrix. Language ladders allow reviewers to report assessment of a factor on a scale of A to D (see an example in Table 2). The primary challenge for language ladder construction was developing reviewer consensus on the definitions of each rung in the ladder. The statements

Table 1
Trainee Evaluation Matrix

The Candidate	Linking Factors	The Research Environment
A1 Academic Record	C1 Linkage to Supervisor's Research	B1 Supervisor's Resources
A2 Research Experience	C2 Role of Trainee	B2 Supervisor's Record
A3 Letters of Reference	C3 Overall Impression of Project	B3 Training Environment

[3] ProGrid® is a registered trademark licensed for use by ProGrid Ventures Inc., Canada

[4] For example, see http://www.innovation.ca/search/viewguide.cfm?guideid=16 and http://www.nserc.ca/pubs/contact/v25_n4_e.pdf.

[5] For example, see http://www.cihr.ca/funding_opportunities/peer_review/peerproc_e.shtml.

Table 2
Academic Record
Language Ladder (Factor A1)

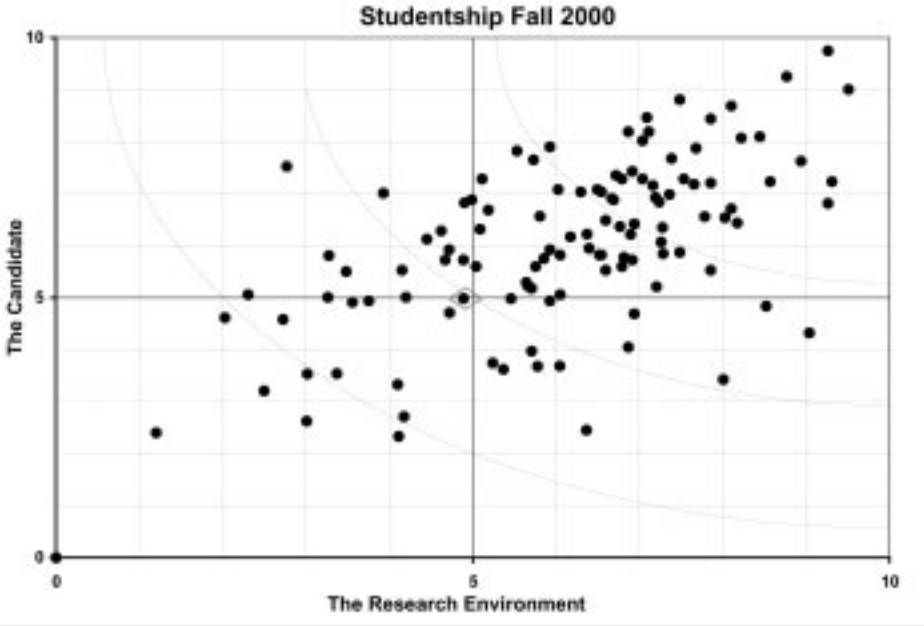
A	Candidate's academic record is adequate for admission to most graduate schools.
B	Candidate has attained above average grades during undergraduate/graduate training and/or candidate has demonstrated steady improvement in grades in the latter stages of training. Candidate received academic recognition (e.g., Dean's List) or a prize or award.
C	Solid, consistently above average academic record throughout the undergraduate/graduate training period and/or candidate demonstrated significant improvement in the academic record in the latter stages of training. Evidence of receipt of several prizes or awards.
D	Outstanding academic record throughout candidate's university level training period. Evidence of receipt of several prizes and awards, some of which are highly competitive, premier awards (e.g., CIHR, NSERC studentships).

defining the rungs were constructed to be distinguishable and as uniformly separated as possible (Bowman, 2001). The statements were constructed from the language used by reviewers in describing their assessments of applications at committee meetings. This is an iterative process between the foundation and the reviewers, and it is expected that the language ladders and performance factors will change over time to reflect the changing values and expectations of the foundation and its stakeholders.

During a competition, reviewers are asked to select their ratings for each of the factors in the *Trainee Evaluation Matrix* based on the information provided by the candidates and their supervisors (i.e., applications and supporting documents such as academic transcripts and letters of reference). Each application is reviewed by three independent reviewers (as opposed to two reviewers in the earlier process).

The reviewers' assessments are then electronically sent to the AHFMR and input into

Figure 1
Example of a ProGrid Output



Note: In the figure, the current applicant is highlighted with a diamond backdrop. Other points on the graph represent other applicants in the same competition.

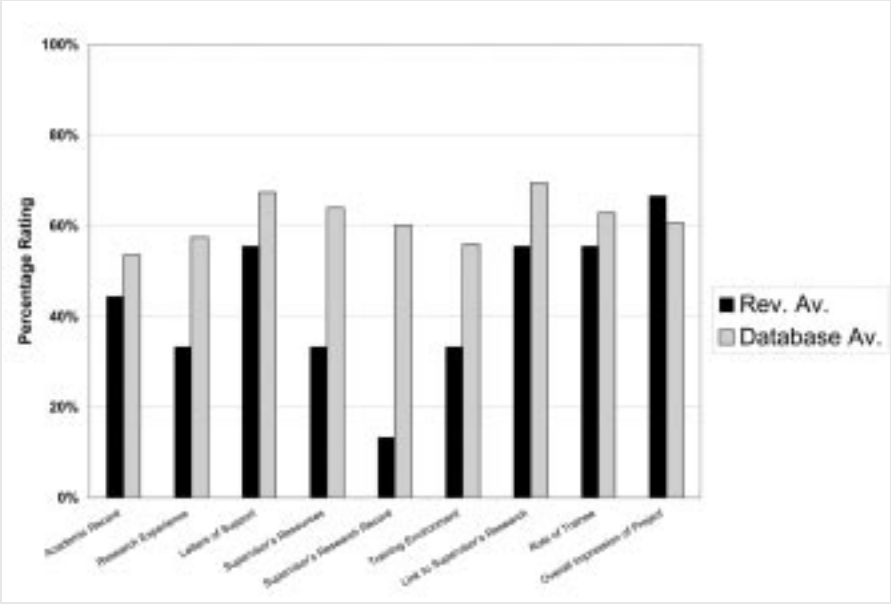
the ProGrid software for analysis. The end result is an automatic output record, in both chart and text form, with the following components: (a) grid position of the application with respect to the two major performance criteria (see Figure 1), (b) specific comments by the reviewers regarding individual criteria or the application as a whole, (c) comparison of the various ratings of each application with the average rating for each performance factor in the overall competition (see Figure 2), (d) customized reports (administrator, reviewers, and candidates), and (e) R value which replaces the rating awarded in the original process.

Added Value

Although the ProGrid-assisted review process is a recent introduction to the AHMFR, some progress made by the AHFMR in its adoption of the software-assisted approach can be clearly demonstrated. The ProGrid software was first used by the AHFMR in its Fall 2000 and Spring 2001 competitions for training awards, which resulted in the distribution of scores shown in Figure 3. While the distributions for the two competitions are quite similar, they are markedly different than the distribution of scores from the more traditional review process. This latter distribution is much more concentrated in the middle (i.e., scores in the 70-79 range), suggesting a higher degree of discrimination for the newer process.

The definition of the process also improved from the foundation's point of view. Whereas the previous system relied primarily on subjective assessments of achievement for each criterion, individual reviewers' weighting of which were largely unknown, the new system uses customized language ladders that more clearly define the meaning of each achievement level (i.e., A, B, C, and D). The review process and scoring is detailed in competition application forms (freely available in the foundation's Web site) and also in the

Figure 2
Example of a Studentship Proposal Profile



Note: In the figure, the applicant's average ratings are shown as "Rev. Av." and the average ratings of all applicants in the competition database are shown as "Database Av."

documentation given to reviewers. This has enhanced the transparency of the review process and allowed for a clearer communication of that same process to all stakeholders.

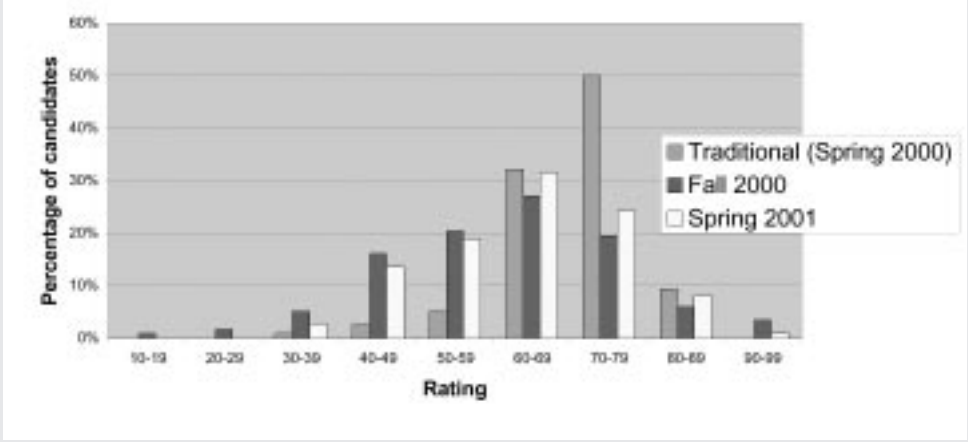
Discussion

The ProGrid database that is created through this process is a rich source of data that can be used in studying relationships between ex ante review and later impact or outcomes of the Foundation's programs. In particular, the capability of the software to provide for longitudinal analysis of candidates relative to the two performance criteria represents potential for examining relative impact of funding on trainees. For example, comparisons of application rating and subsequent performance during the trainee period become possible, as do comparisons of new trainee cohorts with past trainee cohorts. While such studies still require resources, it is anticipated that these will be reduced with the data captured in the ProGrid system.

All of the foregoing is not meant to imply that this new review process is without limitations. For example, experience has shown that when several students from the same lab apply for funding, their training environment (see Table 1) may receive different ratings by different reviewers. This raises the question of the matrix's inter-reviewer reliability. Also, a study of committee scoring at the Heart and Stroke Foundation of Ontario (Hodgson, 1995), found that final scores awarded to applications by review committees were significantly different than scores initially awarded by internal and external reviewers, suggesting that the process of committee discussion is an important element in the peer review process. This clearly has implications for the AHFMR's new process which de-emphasizes committee discussion. Both of these limitations will be studied by the AHFMR, as will others as they arise.

In general, however, the implementation of the new review process for the AHFMR's competitions for training awards has been a positive experience for the foundation and the majority of its stakeholders. While the new process has not been without its critics in the community, it has allowed the AHFMR to address specific process issues affecting the candidates, the reviewers, and the foundation itself, and has increased the organization's internal capacity for self-evaluation. The pilot project described in this paper was applied first to the studentship application review process. This approach has now been extended to other programs (e.g., summer students, fellows, clinical fellows) and a more general

Figure 3
Ratings Comparison



use of this approach for program evaluation and impact analysis is also being explored.

Conclusions

The new process allows the foundation's administrators to evaluate the usage of specific performance factors by reviewers to confirm the consistency of their approach and the acceptability of factors. This allows for critical examination of the review process and has allowed administrators to merge a number of performance factors where the use of individual factors either added little to the process or were clearly linked to each other (e.g., the factors linkage to supervisor's research and role of trainee were merged into one criterion, role of trainee and linkage to supervisor's research). These changes were made on the basis of reviewer and applicant feedback.

The software also supports comparisons of decision-making patterns between committees. For example, the AHFMR has a Health Trainee Advisory Committee (concerned with applications in the fields of population health, behavioral research, and other non-biomedical health sciences), a Studentship Advisory Committee (concerned with applications in the fields of the biomedical sciences), and two Fellowship Advisory Committees. The decision- making patterns of these committees bear examination and comparison, something that would have been difficult under the old system but may be simplified under the new system.

Foundation administrators find the new process to be timely and possibly more cost-effective than the original review process. Committees either meet only to obtain information about the outcome of their work or to discuss more a limited number of outlier proposals. Although technically this had been possible for a long time, procedurally, the need to discuss each application as part of the original process precluded much pre-meeting triaging within given competitions.

Also, the responsiveness of the competitions to the information needs of the applicants improved. Providing applicants with additional and more meaningful feedback was a stated goal of the process and the applicant summary report generated by ProGrid is welcomed by most applicants (although there are still improvements that can and are being made to the system).

Finally, the new process makes different use of committee members' time at meetings (two per year) for all competitions. While the original process required that a large percentage of time of in-person meetings be devoted to discussion of all of the applications, meetings can now be devoted to discussing those applications where there is a significant difference of opinion between committee members, with the remainder of the time spent discussing competition policy, as recognized through the pre-meeting assessments provided to the AHFMR. The process also enables shorter meetings (and possibly the elimination of review meetings altogether), which leads to a better use of scarce reviewer resources. The limited experience that the Foundation has with the decision-assist tool also suggests that even in cases where there is disagreement between reviewers, the committee discussion rarely results in a significantly different outcome for the applicants. Most proposal ratings change no more than five percent.

In addition to these outcomes, the introduction of the ProGrid-facilitated process has other implications for program evaluation and review at the AHFMR. As mentioned, it allows administrators and staff to perform sophisticated comparisons between competitions, groups of reviewers, and subgroups of applicants. However, there is also the potential to extend this approach beyond ex ante review to ex post program evaluation to program outcomes.

References

Bowman, C.W. (2001). Evaluating intellectual capital—III. Evaluating investment opportunities. *Canadian Chemical News,* 53, 30-32.

Brzustowski, T. (2000a). Is peer-review fatigue setting in? *NSERC Contact,* 25, 1-2.

Brzustowski, T. (2000b). Peer-review fatigue and the need to double

NSERC's budget. *NSERC Contact,* 25, 1-2.

Cunningham, P., Boden, M., Glynn, S., & Hills, P. (2001). *Measuring excellence in government science and technology: International practices: France, Germany, Sweden and the United Kingdom*. Ottawa, ON: S&T Strategy Directorate, Industry Canada.

Geisler, E. (2000). *The metrics of science and technology.* Westport, CT: Quorum Books.

Hodgson, C. (1995). Evaluation of cardiovascular grant-in-aid applications by peer review: influence of internal and external reviewers and committees. *Canadian Journal of Cardiology,* 11, 864-868.

Kostoff, R. N. (1992). Federal research impact assessment methods. *Research Management Review,* 6, 22-44.

Luukkonen-Gronow, T. (1987). Scientific research evaluation: A review of methods and various contexts of their application. *R&D Management,* 17, 207-221.

NIAID Council (1999). Understanding percentiles and other funding factors. *NIAID Council News,* 8, 6.

Smith, W. (2001). *Measuring excellence in government science and technology: International practices: New Zealand and Australia*. Ottawa, ON: S&T Strategy Directorate, Industry Canada.

United States General Accounting Office (1997). *Measuring performance: Strengths and limitations of research indicators* (Rep. No. GAO/RCED-97-91). Washington, DC: United States General Accounting.