International Journal of Surgery Open 4 (2016) 38–41

Contents lists available at ScienceDirect

International Journal of Surgery Open

journal homepage: www.elsevier.com/locate/ijso

Research Paper

Comparative outcomes of face-to-face and virtual review meetings

Nghia M. Vo *, Gabrielle M. Quiggle, Kishena Wadhwani

Division of Scientific Review, Office of Extramural Research, Education and Priority Population, Agency of Healthcare Research and Quality, Rockville, MD, USA

ARTICLE INFO

Article history:
Received 24 May 2016
Received in revised form 29 July 2016
Accepted 30 July 2016
Available online 3 August 2016

Keywords:
Peer review
Face-to-face meeting
Virtual review
Comparative outcome
Grant application and funding

ABSTRACT

Background: The present study proposes to evaluate the effects of face-to-face (FF) and virtual review (VR) sessions on peer reviewers' scores and consistency of peer review.
Methods: Retrieved review sessions conducted between 2012 and 2014 yielded 119 and 51 discussed applications for the FF and VR groups, respectively. Changes between preliminary scores, post discussion scores and final matrix scores were analyzed. Consistency between the two meeting modalities was measured by percentage and increments of score changes.
Results: Discussion changed the preliminary scores in 37% of applications reviewed in the FF group and 24% of applications reviewed in the VR group (no difference between groups). Applications that received a preliminary score in the 10 to 30 point-range were more positively than negatively impacted by discussion in both modalities. FF discussion led to a wider range of scoring changes (−10 points to 17 points) than VR (−7 points to 10 points), but discussion was not found to differentially improve or worsen scores between the two modalities. When comparing post-discussion and final matrix scores, 27 (23%) applications' scores changed in the FF meetings compared to 13 (25%) in the VR meetings (no difference between groups).
Conclusions: FF and VR sessions result in (1) minimal differences in preliminary scores, (2) nonsignificant percentage changes in scoring, and (3) non-significant change in the percentage of magnitude of scoring. The two review methods appear to be similar in evaluating grant applications.

1. Introduction

Peer review of grant applications can be done through either face-to-face (FF) or virtual review (VR) meetings. The latter is a web-assisted technology that allows reviewers to communicate with each other through a web-based platform. The Agency for Healthcare Research and Quality (AHRQ) conducts VR meetings using the WebEx system, which has audio, high definition 2 × 2 video, real-time content sharing, and the capability feed for up to seven simultaneous webcam videos [1]. Although VR has been performed hundreds of times in the past, only one study from the American Institute of Biological Sciences has carefully evaluated the effect of FF and VR meetings on the peer review of grant applications [2]. That study used the old 5-point scoring system (from 1 to 5) where 1 is the best score and 5 is the worst score.

The present study proposes to evaluate the effects of FF and VR meeting modalities on reviewers' scores using the new 9-point scoring system (from 1 to 9) where 1 is the best and 9 is the worst score [3]. This study will add to the understanding of the advantages/disadvantages of each review system and increase the knowledge of what and how reviewers think and react to the discussions in each particular setting.

2. Methods

The Division of Scientific Review (DSR) reviews all applications submitted to AHRQ in response to Funding Opportunity Announcements (FOA). An FOA could be a Program Announcement (PA), which occurs three times each year, or a Request for Application (RFA), which is a specific one-time request. The DSR has five study sections aligned with a particular portfolio: Healthcare Information Technology Research (HITR), Healthcare Patient Safety and Quality Improvement Research (HSQR), Healthcare Research and Training (HCRT), Healthcare Systems and Value Research (HSVR), and Healthcare Effectiveness and Outcomes Research (HEOR). The Special Emphasis Panel (SEP), on the other hand, reviews RFAs and may address some program portfolios.

The grant application review process goes through two stages. Each application is first assigned to three reviewers who evaluate the scientific merit of the application and provide a preliminary impact or pre-discussion score. Second, when the review meeting

Part of this paper was presented at the 2014 American Public Health Association (APHA) Annual Meeting, New Orleans, LA, on November 18, 2014.

* Corresponding author. Division of Scientific Review, Office of Extramural Research, Education and Priority Population, Agency for Healthcare Research and Quality, 5600 Fishers Lane, Rockville, MD 20857, USA. Fax: 301 427 1562.
E-mail address: nghia.vo@ahrq.hhs.gov (N.M. Vo).

http://dx.doi.org/10.1016/j.ijso.2016.07.002

convenes, the application is discussed by the three lead reviewers with the input of the rest of the panelists, followed by the lead reviewers restating their own scores. These scores are known as lead reviewers' post-discussion scores. The remaining panelists then write down their own scores. The average of all the reviewers' scores is multiplied by a factor of 10 to get rid of decimals. The final product is known as the matrix or final impact score and this score determines the ranking of the application reviewed. Applications that receive a matrix score between 10 and 30 points have a high likelihood (estimated at 50–60%) of being funded.

Not all the submitted applications are discussed. To allow more time to discuss the meritorious applications, about half are "triaged" and not discussed [4]. Since these applications do not receive post discussion scores, they are not included in this study.

The purpose of funding research grants is to "fund the best science, by the best scientists" and "to see that NIH [and AHRQ] grant applications receive fair, independent, expert, and timely reviews—free from inappropriate influences—so that NIH [and AHRQ] can fund the most promising research." [5–7] The purpose of the peer review is to identify the best of these applications. Over time, federal institutions have modified components of the review process to adapt to new demands and changes. Thus, the old 5-point scoring system [2,8] was replaced in October 2009 by a 9-point system [3] to give reviewers the chance to spread out their scores.

Eleven review meetings conducted by one scientific review officer (SRO) between 2012 and early 2014 were retrieved for analysis. The VR meetings were typically one-day sessions, although one was conducted as a 2-day session and involved 34 applications. The data analyzed were broken down into (1) preliminary or pre-discussion scores, (2) post-discussion scores, (3) average final or matrix scores, (4) magnitude of differences between preliminary and post-discussion-scores, and (5) magnitude of differences between post-discussion scores and matrix scores. To be consistent throughout the study, the average preliminary, post-discussion, and final matrix scores were multiplied by a factor of 10 to get rid of decimals. For some analyses, scores were categorized into the following four impact levels: 10–20, 21–30, 31–40, and >40.

Consistency of the review was measured in terms of percentage and increments of score changes.

Given the non-normal distribution of the data, Fisher's exact test was conducted to compare categorical data between FF and VR meetings. Mann–Whitney *U* tests and Wilcoxon sign-ranked tests were also calculated to determine if the medians of preliminary, post discussion, and final matrix scores statistically differed between and within each meeting modality.

3. Results

Data from six FF sessions and five VR meetings were collected for analysis totaling to 119 (FF) and 51 (VR) discussed applications.

3.1. *Effect of discussion on lead reviewers' preliminary scores*

One third of the discussed applications (n = 56, 33%) had their average preliminary impact scores change score categories after discussion. Of these, changes for better occurred in 21% of the applications (n = 12) and for worse in the remaining 79% (n = 44). When analyzed separately, 44 (37%) applications' scores changed (n = 34 worse, n = 10 better) in the FF meetings (Table 1) and 12 (24%) changed (n = 10 worse, n = 2 better) in the VR meetings (Table 2). The proportion of applications that changed impact levels after discussion was found not significantly different between the two FF and VR (p = 0.109). Applications that received a preliminary score in the 10 to 30-point range were more positively than negatively impacted by discussion in both modalities (p = 0.369): n = 40 (66% of 61 applications) improved or had no meaningful change in the FF group; n = 29 (76% of 38 applications) improved or had no meaningful change in the VR group.

Table 1
Change from preliminary to post-discussion impact score categories, face-to-face.

Preliminary scores	Post-discussion scores				
	10–20	21–30	31–40	41–90	Total
10–20	10 (8%)	2 (2%)	–	–	12 (10%)
21–30	5 (4%)	25 (21%)	16 (13%)	3 (3%)	49 (41%)
31–40	–	4 (3%)	18 (15%)	13 (11%)	35 (29%)
41–90	–	–	1 (1%)	22 (18%)	23 (19%)
Total	15 (13%)	31 (26%)	35 (29%)	38 (32%)	119 (100%)

Shadings highlight the concordance of pre- and post-discussion scores.

Table 2
Change from preliminary to post-discussion impact score categories, virtual review.

Preliminary scores	Post-discussion scores				
	10–20	21–30	31–40	41–90	Total
10–20	12 (24%)	5 (10%)	–	–	17 (33%)
21–30	2 (4%)	15 (29%)	4 (8%)	–	21 (41%)
31–40	–	–	11 (22%)	1 (2%)	12 (24%)
41–90	–	–	–	1 (2%)	1 (2%)
Total	14 (27%)	20 (39%)	15 (29%)	2 (4%)	51 (100%)

Shadings highlight the concordance of pre- and post-discussion scores.

The impact of discussion on the magnitude and direction of change in scores was also assessed using the raw preliminary and post discussion scores (Tables 3 and 4). Overall, FF discussion led to a wider range of scoring changes (–10 points to 17 points) than VR (–7 points to 10 points). There was a statistically significant difference in the proportions of changed scores between FF (n = 86, 72%) and VR (n = 26, 51%) modalities (p = 0.009) with more score changes occurring in the FF, but discussion was not found to differentially improve or worsen scores between the two modalities (FF: n = 24 better, n = 62 worse; VR: n = 6 better, n = 20 worse) (p = 0.801).

A Wilcoxon signed-rank test showed that discussion elicited a statistically significant change in post-discussion scores in both FF and VR modalities (Z = –4.08, p < 0.001; Z = –2.51, p = 0.012,

Table 3
Change from preliminary to post-discussion scores, face-to-face.

	Preliminary scores					
Changes	10–20	21–30	31–40	41–90	Total	
–11 to –20	–	–	–	–	–	Better
–6 to –10	2 (2%)	3 (3%)	4 (3%)	3 (3%)	12 (10%)	
–1 to –5	1 (1%)	5 (4%)	4 (3%)	2 (2%)	12 (10%)	
0	7 (6%)	16 (13%)	6 (5%)	4 (3%)	33 (28%)	
1 to 5	1 (1%)	8 (7%)	9 (8%)	7 (6%)	25 (21%)	Worse
6 to 10	1 (1%)	13 (11%)	10 (8%)	5 (4%)	29 (24%)	
11 to 20	–	4 (3%)	2 (2%)	2 (2%)	8 (7%)	
Total	12 (10%)	49 (41%)	35 (29%)	23 (19%)	119 (100%)	

Table 4
Change from preliminary to post-discussion scores, virtual review.

	Preliminary scores					
Changes	10–20	21–30	31–40	41–90	Total	
–11 to –20	–	–	–	–	–	Better
–6 to –10	–	1 (2%)	–	–	1 (2%)	
–1 to –5	2 (4%)	3 (6%)	–	–	5 (10%)	
0	8 (16%)	10 (20%)	7 (14%)	0 (0%)	25 (49%)	
1 to 5	6 (12%)	4 (8%)	4 (8%)	–	14 (27%)	Worse
6 to 10	1 (2%)	3 (6%)	1 (2%)	1 (2%)	6 (12%)	
11 to 20	–	–	–	–	–	
Total	17 (33%)	21 (41%)	12 (24%)	1 (2%)	51 (100%)	