Quality Control of Epidemiological Lectures Online: Scientific Evaluation of Peer Review

Faina Linkov, Mita Lovalekar, Ronald LaPorte

Department of Epidemiology,
Graduate School of Public Health,
University of Pittsburgh, Pa, USA

Aim To examine the feasibility of using peer review for the quality control of online materials.

Methods We analyzed the inter-rater agreement on the quality of epidemiological lectures online, based on the Global Health Network Supercourse lecture library. We examined the agreement among reviewers by looking at κ statistics and intraclass correlations. Seven expert reviewers examined and rated a random sample of 100 Supercourse lectures. Their reviews were compared with the reviews of the lay Supercourse reviewers.

Results Both expert and non-expert reviewers rated lectures very highly, with a mean overall score of 4 out of 5. Kappa (κ) statistic and intraclass correlations indicated that inter-rater agreement for experts and non-experts was surprisingly low (below 0.4).

Conclusions To our knowledge, this was the first time that poor inter-rater agreement was demonstrated for the Internet lectures. Future research studies need to evaluate the alternatives to the peer review system, especially for online materials.

> **Correspondence to:**

Faina Linkov
Department of Medicine
University of Pittsburgh
3512 Fifth Avenue, Room 312
Pittsburgh, PA 15261, USA
fyL1@pitt.edu

> **Received:** January 8, 2007
> **Accepted:** January 17, 2007

> **Croat Med J. 2007;48:249-55**

The South Korean research scandal (1) not only had a major effect on stem cell research but also the peer review process itself. In a recent New York Times article, Dr Kennedy, the editor of Science, indicated that "peer review is not a process that guarantees truth" (2). In this report, we take a scientific approach to examine the process of peer review. Perhaps the best definition and function of peer review is given in the Wikipedia, the open source encyclopedia: "Peer review is a process of subjecting an author's scholarly work or ideas to the scrutiny of others who are experts in the field. It is used primarily by publishers, to select and to screen submitted manuscripts, and by funding agencies, to decide the awarding of monies for research." Peer review is being recognized as one of the ways to control the quality of biomedical publications.

Although the beginnings of "peer review" are frequently associated with the Royal Society of London when it took over official responsibility for the Philosophical Transactions in 1752, antecedents of peer review practices go back to the 17th century (3). In the past three centuries, peer review has been viewed almost as an extension of the scientific method itself – a gold standard. Despite such a long history in science, recent articles suggest that the process of peer review may be in crisis (4) and may need to undergo some significant changes. A systematic review of the biomedical literature conducted by Jefferson et al (5) concluded that there is very little science behind the peer review process. Perhaps the major problem of peer review is that, although it is thought to be a gold standard of science and utilized extensively for selecting articles for publication, grants for funding, and abstracts for presentations at conferences, it has not gone under scientific scrutiny. By now, there is a widespread consciousness among scientists regarding the deficiencies of peer review (6). It is slow, expensive, subjective, prone to bias, and easily abused (7). Another drawback of this system is the tendency to select against novel work (8). Procedures that are currently used by many professional journals, such as blind or masked review, may not completely alleviate the effects of peer review pitfalls (9), and interestingly, peer review practices vary from journal to journal (10). The research needed to understand the broader effects of peer review poses many methodological problems and would require the cooperation of many parts of the scientific community (11).

We investigated peer review in a rising area of research communication, that of the web-based scientific research communication in the form of lectures. There have been a few studies investigating the science of journal and grant peer review (5), but there have been even fewer investigating peer review of web lectures. Mechanisms of monitoring the quality of lectures is becoming more and more important, as over the past 2 years the number of PowerPoint lectures on the web has increased from 5 to 25 million files. If an instructor wants to utilize an existing lecture on the Internet for his or her course, there must be a mechanism for them to find out if such lecture is valid, trustworthy, and updated.

The Supercourse (12-14) is a library of over 3000 epidemiological and public health lectures (as of January 2007), targeting the educators. It is a project based in the University of Pittsburgh, Department of Epidemiology, and supported by the National Library of Medicine of the National Institutes of Health (NIH). Over 40000 researchers from 151 countries of the world are working together to share their best lectures in the area of epidemiology and public health in the Supercourse. Supercourse's aim is to improve global research and training by sharing high quality lectures. In this research study, we evaluated the agreement among Supercourse reviewers, using a web-based system for peer review of the quality of information contained in the lecture. Each Supercourse lecture consists of 14 to 32 consecutive pages and every page has a uniform format: a slide with 320 by 240 pixels in size on the left and text beside the slide on the right. On

the last page of each lecture, there is the peer review form for the lecture. This page allows the readers of the lecture to rate and give comments on the lecture. Review forms of the Supercourse lectures became the basis of quality control for the Supercourse lectures and the data collected through these forms helped us to test our hypothesis.

The main goal of our study of peer review was to address the question: What is the reliability of peer reviewers? If one reviewer rates research communications as excellent and another rates the same module as poor, this would throw into question the ability of a peer review system to select the optimal lectures for educating students. Although several articles have been published on reliability of peer reviewers for traditional research communications, this is one of the first papers to evaluate the reliability of scientists reviewing online research lectures.

In this paper, we make no assumption that peer reviewers have to agree on the quality of a paper. In fact, there is usually no requirement that the referees achieve consensus from the editor's point of view (*http://en.wikipedia.org/wiki/Peer_review*). However, in general, the paper has a very good chance of being published if all reviewers agree that the paper is good. On the other hand, the chances of publishing are very slim if all reviewers suggest that it should be rejected. Similar situation can be seen with peer review of grants and conference abstracts, which is why it is important to look at the agreement among peer reviewers.

Methods

Review form

The lecture review forms rate the lecture on content, presentation, relevance, and provide an overall rating, using a 5-point Likert scale (5 – excellent, 4 – above average, 3 – average, 2 – below average, 1 – poor). Lecture review form also has a text box for any additional comments and/or feedback. Lecture review form differs from the form regularly utilized by biomedical journals, because it utilizes Likert-type scales to reduce the time needed for lecture evaluation. Additionally, since lectures do not follow the traditional format of research articles (introduction, methods, results, discussion), the form utilized for lecture review had to have a different structure.

Data collection

To examine the consistency of peer reviewers, we selected a random sample of 100 Supercourse lectures using computer generated random numbers. Only lectures that previously accumulated three or more reviews were selected for the study. It was decided to concentrate our evaluation efforts on the first 1000 lectures, since these were the ones that accumulated the maximum number of reviews. Only lectures in English language were considered in this study. All 100 lectures and instructions were accessible to study participants through a Web site.

Study participants

Seven expert peer reviewers from 6 countries (USA, France, Cuba, UK, China, India) agreed to participate in the project. The main criteria for eligibility to participate in this study were MD or PhD degree and experience with being a peer reviewer. Three of the reviewers who participated in this study were journal editors. Approximately 40% of the reviewers who were approached for this study agreed to participate. These were likely a representative sample of global biomedical reviewers, as all of them were experienced in peer review process. Most of our experts were MDs and PhDs with extensive record of peer reviewed publications. The qualifications of reviewers were similar to the pool of reviewers commonly used by the major biomedical journals. Each reviewer has been assigned a number from 1-7, under which the data were analyzed.

Data collection tool and outcome measures

We collected multiple ratings for each lecture, including presentation and relevance. The peer review form used in this study is the same form that is utilized for all Supercourse lectures (*http://www.pitt.edu/~super1/lecture/lec10511/review.htm*). The question 13 of the review form asks reviewers to give the "overall score" for the lecture. For the purpose of this paper we only concentrated on the "overall score" parameter, as we were interested in evaluating the overall quality of the lecture, not the other parameters assessed by the lecture review form.

Statistical analysis

Kappa statistics were calculated in order to examine the agreement among the reviewers. Intraclass correlations were also calculated to analyze the similarities among the ratings. Existing literature in the area suggests that intraclass correlation is commonly used to measure inter-rater agreement. Basic descriptive statistics were calculated to analyze basic lecture review trends of expert and non-expert reviewers in the Supercourse. Statistical analysis were performed using SAS software (SAS Institute, Inc., Cary, NC, USA)

Intraclass correlation is large and positive when there is no variation within the groups, but group means differ. It will be at its largest negative value when group means are the same but there is great variation within groups. Its maximum value is 1.0, but its maximum negative value is [-1/(n-1)]. A negative intraclass correlation is not common, but it occurred in our study. Negative intraclass correlations occur when between-group variation is lower than within-group variation, indicating that a third (control) variable has introduced non-random effects on the different groups.

Results

All the data were collected over the period of 2 months. Each reviewer was assigned 100 lectures to review. A total of 658 lecture reviews were collected from 7 experts, indicating that all experts reviewed over 90% of lectures that were assigned to them. We do not know why reviewers did not review some lectures, as each reviewer omitted different set of lectures, but we suspect that some were omitted due to lack of interest or lack of time. Overall, the lectures were reviewed positively by the experts, with the mean overall score of 3.92/5 and standard deviation of 0.95. Non experts also gave very positive reviews (mean ± statistical deviation [SD] score was 4.12 ± 0.82).

Intraclass correlations and Kappa statistics were calculated to examine the inter-rater agreement among experts and the Supercourse reviewers (non-experts). Kappa value ranges from 0 (poor or no agreement) to 1 (perfect agreement). If peer review had high inter-rater agreement, we would expect to see high correlations and kappa statistics (close to 1). There was no relationship among any of the 7 expert reviewers, despite their high level of expertise (Table 1). Although presented data concentrates on the "overall score" parameter, we also found very poor agreement for other parameters, such as relevance and presentation (data not shown).

Intraclass correlations were calculated to examine the inter-rater agreement between experts and the Supercourse reviewers (non-experts).

Table 1. Inter-rater agreement: Kappa statistics for the "overall quality of the lecture" rating among seven expert reviewers participating in the study

	Reviewer No.						
Reviewer No.	1 (n = 94)	2 (n = 103)	3 (n = 81)	4 (n = 99)	5 (n = 97)	6 (n = 94)	7 (n = 91)
1 (n = 94)	1	0.04	-0.03	0.06	-0.04	0.05	0.03
2 (n = 103)		1	0.04	0.02	0.01	0.02	-0.04
3 (n = 81)			1	-0.06	0.04	-0.01	0.04
4 (n = 99)				1	0.13	-0.05	-0.01
5 (n = 97)					1	0.12	0.01
6 (n = 94)						1	0.12
7 (n = 91)							1

Table 2. Inter rater-agreement: intra class correlation coefficients for the "overall quality of the lecture rating" among seven expert reviewers and Supercourse (non-expert) reviewers

Reviewer number	1 (n = 94)	2 (n = 103)	3 (n = 81)	4 (n = 99)	5 (n = 97)	6 (n = 94)	7 (n = 91)	Supercourse (n = 849)
1 (n = 94)	1	0.49	-0.25	-0.45	-0.40	-0.43	0.07	-0.28
2 (n = 103)		1	0.31	0.12	0.19	0.03	0.24	-0.31
3 (n = 81)			1	0.17	0.14	0.12	0.12	-0.26
4 (n = 99)				1	-0.18	-0.33	-0.33	-0.17
5 (n = 97)					1	-0.33	-0.45	-0.38
6 (n = 94)						1	-0.84	-0.11
7 (n = 91)							1	-0.17
Supercourse (n = 849)*								1

*Supercourse reviewer data refer to "non-expert" reviewers visiting Supercourse sites. All Supercourse users are given a chance to review any lecture they wish. This formed a foundation of lecture quality control in the Supercourse.

The data suggest that experts' reviews, as well as the Supercourse (non-expert) reviews, poorly correlate with one another (Table 2).

Discussion

Our study showed that both expert and non-expert reviewers rated lectures very highly. We performed Kappa analysis to see whether reviewers agree in their judgments on specific lectures. Since Kappa is a standard measure for evaluating agreement, we wanted to utilize this method, so that we could compare the results of our study to other investigations in the field of peer review evaluation. As in the existing literature on peer review (15), Kappa statistic and intraclass correlations indicated that inter-rater agreement for experts and non-experts was very low.

There have been very few scientific studies examining peer review. Our study is, to the best of our knowledge, one of the first efforts to apply scientific method to evaluate the consistency of reviewer's judgments of quality for Internet-based lectures. What can the results of this study mean for the peer review system as a whole? A study similar to ours investigated the agreement between two referees in the evaluation, based on a 4-point scale checklist of abstracts submitted for a primary care conference. The Kappa statistic for inter rater agreement on subjective questions like "importance" ranged from 0.01 to 0.25 (16), which is similar to our results. The agreement among peer reviewers was also analyzed in the *Croatian Medical Journal* and the *Lancet*, with poor to fair Kappa statistic for both national and international articles (17). Poor agreement among reviewers was also found for grant reviewers (15,18).

Lack of agreement among peer reviewers, demonstrated in our study and other existing literature sources, indicates that peer review system may not be good for identifying top quality intellectual property. One may argue that the inter-rater agreement in our study was low just because the reviewers were not properly trained to review the materials. However, this is unlikely because the expert reviewers who were selected represented a highly experienced group. Moreover, few if any reviewers are trained to review articles or grants. The literature in this area suggests that even if you train a reviewer in a group session to do a better job at peer review, there is only a slight impact on the quality of peer review (19); some studies even suggest that additional training has absolutely no affect on the quality of the review (20,21). If there is no consistency even among highly trained peer reviewers, then the effectiveness of the whole system is in question.

Our study had several limitations. First of all, the review form that was used for the Supercourse lectures is slightly different than the standards utilized for reviewing articles in biomedical journals. If we used the same format as journals, our results would probably be more generalizable. Additionally, some of the Supercourse lectures that were included in this research had

grammar mistakes. We could see that such lectures received low scores from some reviewers, but not the others, potentially leading to lower inter-reviewer agreement.

The system of peer review cannot consistently identify high quality materials, as opposed to materials of poor quality. In the future studies of peer review, we may want to concentrate on evaluating a single aspect of a lecture and see if there is an agreement there. As scientists, we cannot expect peer review to provide solution to the problem of quality control of scientific communications when "peers" do not agree in their quality judgments. This study helped to emphasize that peer review system is a tool that may be flawed. Just like any tool utilized by scientists, peer review has to undergo the scrutiny of the scientific method in order to test its effectiveness. Donald Kennedy, said replication "is the ultimate test of truth in science" (2). Although true replication is not always possible in laboratory research, the Internet may be an important medium that can accommodate the publication of multiple studies in the same area. For example, Supercourse library has multiple lectures on the same topic, such as molecular epidemiology and breast cancer research. Replication of publications and lectures, as demonstrated in the Supercourse, is very important for the future of peer review. One might argue that peer reviewing a lecture is different than peer reviewing an article, but in reality they are not that different. Large scale studies are equally needed for the millions of scientific articles, as well as for the burgeoning number of scientific lectures. Further research in this area needs to be done.

We use peer review approach for journal articles, conference abstracts, grants, and for the lectures. Perhaps all of these applications are judged by the method of no proved validity. What is important is not to take sides as to whether peer review works or not but rather to scientifically evaluate peer review and other forms of quality control. It is important to point out that quality control does work in other areas, especially in industry (3). The existing quality control systems for the Internet based scientific materials are not optimal. They need to be researched and further developed so that scientists around the globe could trust the Internet-based information. A change in the process of peer review will not be possible without challenging traditional paradigms and exploring new alternatives, such as consumer driven quality control mechanisms. Future research of quality control of the biomedical publications could concentrate on the use of Deming-like systems of total quality management, successfully utilized in industry (22). William Edwards Deming was an American statistician, widely credited with improving production in the United States during World War II using statistical quality control systems. He is perhaps best known for his work in Japan, where from 1950 onward he taught top management how to improve design (and thus service), product quality, testing, and sales (22). This research made it clear that the scientific community, especially in the area of medicine, is in need of an improved science of quality control. Implementation of new quality control mechanisms for biomedical literature and web materials will need to engage all the stakeholders involved in this process. Our ultimate goal should be to bring the success of quality control in the industry to biomedical journals.

Acknowledgment

This project has been carried out with the support of the Global Health Network Supercourse Project, housed at the Department of Epidemiology, University of Pittsburgh. We would like to thank Weiting Xie for her help with the data collection and statistical analysis. Source of support for this study was Epidemiology on the Internet grant by NLM, NIH.

References

1 Steinbrook R. Egg donation and human embryonic stem-cell research. N Engl J Med. 2006;354:324-6. Medline:16436762

2 Dean C. Where science and public policy intersect, researchers offer a short lesson on basics. The New York Times. January 21, 2006. Available from: *http://www.*

nytimes.com/2006/01/31/science/31cong.html?ex=1296363600&en=73143d5abbac317b&ei=5090&partner=rssuserland&emc=rss. Accessed: March 12, 2007.

3 Kronick DA. Peer review in 18th-century scientific journalism. JAMA. 1990;263:1321-2. Medline:2406469

4 Mulligan A. Is peer review in crisis? Oral Oncol. 2005;41:135-41. Medline:15695114

5 Jefferson T, Alderson P, Wager E, Davidoff F. Effects of editorial peer review: a systematic review. JAMA. 2002;287:2784-6. Medline:12038911

6 Ingelfinger FJ. Peer review in biomedical publication. Am J Med. 1974;56:686-92. Medline:4825604

7 Thorn A. Peer review: a closed system in need of reform [In Swedish]. Lakartidningen. 2002;99:3106-8. Medline:12198928

8 Olson CM. Peer review of the biomedical literature. Am J Emerg Med. 1990;8:356-8. Medline:2194471

9 Hojat M, Gonnella JS, Caelleigh AS. Impartial judgment by the "gatekeepers" of science: fallibility and accountability in the peer review process. Adv Health Sci Educ Theory Pract. 2003;8:75-96. Medline:12652170

10 Good CD, Parente ST, Rennie D, Fletcher SW. A worldwide assessment of medical journal editors' practices and needs – results of a survey by the World Association of Medical Editors. S Afr Med J. 1999;89:397-401. Medline:10341824

11 Jefferson T, Wager E, Davidoff F. Measuring the quality of editorial peer review. JAMA. 2002;287:2786-90. Medline:12038912

12 LaPorte RE, Linkov F, Villasenor T, Sauer F, Gamboa C, Lovalekar M, et al. Papyrus to PowerPoint (P 2 P): metamorphosis of scientific communication. BMJ. 2002;325:1478-81. Medline:12493674

13 Laporte RE, Omenn GS, Serageldin I, Cerf VG, Linkov F. A scientific supercourse. Science. 2006;312:526. Medline:16645075

14 Linkov F, LaPorte R, Lovalekar M, Dodani S. Web quality control for lectures: Supercourse and Amazon.com. Croat Med J. 2005;46:875-8. Medline:16342339

15 Hodgson C. How reliable is peer review? An examination of operating grant proposals simultaneously submitted to two similar peer review systems. J Clin Epidemiol. 1997;50:1189-95. Medline:9393374

16 Montgomery AA, Graham A, Evans PH, Fahey T. Inter-rater agreement in the scoring of abstracts submitted to a primary care research conference. BMC Health Serv Res. 2002;2:8. Medline:11914164

17 Marusic A, Mestrovic T, Petrovecki M, Marusic M. Peer review in the Croatian Medical Journal from 1992 to 1996. Croat Med J. 1998;39:3-9. Medline:9475799

18 Cole S, Cole JR, Simon GA. Chance and consensus in peer review. Science. 1981;214:881-6. Medline:7302566

19 Schroter S, Black N, Evans S, Carpenter J, Godlee F, Smith R. Effects of training on quality of peer review: randomised controlled trial. BMJ. 2004;328:673. Medline:14996698

20 Callaham ML, Wears RL, Waeckerle JF. Effect of attendance at a training session on peer reviewer quality and performance. Ann Emerg Med. 1998;32:318-22. Medline:9737493

21 Callaham ML, Knopp RK, Gallagher EJ. Effect of written feedback by editors on quality of reviews: two randomized trials. JAMA. 2002;287:2781-3. Medline:12038910

22 Edwards Deming W. Available from: *http://en.wikipedia.org/wiki/W._Edwards_Deming*. Accessed: March 13, 2007.