A Study on the Effectiveness of Peer Review Meeting

Noriyoshi Kuno Tsuyoshi Nakajima
Design Systems Engineering Center
Mitsubishi Electric Cooperation
Kamakura, Japan
Kuno.Noriyoshi@cb.MitsubishiElectric.Co.jp

Makoto Matsushita Katsuro Inoue
Department of Computer Science
Graduate School of Information Science and Technology
Osaka University
Osaka, Japan

***Abstract*— The effectiveness of peer review meetings in software development has been discussed for many years. Porter concludes that peer review meetings do not contribute significantly to defect extraction. This paper shows contradictory data to the Porter's findings and our interpretation for them.**

Keywords—inspection; peerreview; meetings

I. INTRODUCTION

Social system failures caused by software defects have been increased. Therefore it is crucially important to establish appropriate quality control methods and technologies for software development. Peer review is one of the most widely used methods to verify artifacts at each development phase, through requirements analysis to coding.

It is widely known that the effectiveness of peer reviews varies to large degree. To reduce the variability, a lot of improved variations for peer reviews have been proposed, such as checklist-based and perceptive-based reading.

Peer review meeting is a main part of peer review activities in the original peer review methods [1]. However, the effectiveness of them had been contentious for a long time [2][3]. Porter assessed the effectiveness for peer review meetings by analyzing data from peer reviews with peer review meetings and without a peer review meeting [4][5]. Porter concludes that peer review meetings do not contribute significantly to detect extraction.

We decided to examine the effectiveness of the peer review meeting because peer review meetings are widely used in the field of software development despite Porter's findings,

In this paper, we will present findings that come to an opposite conclusion from Porter's, i.e., peer review meetings can be sufficiently effective in the real situations using a controlled manner.

II. EFFECTIVENESS OF A PEER REVIEW MEETINGS

A. Peer review process

Fig.1 shows the main process of peer reviews to be discussed, which has three parts: individual check, peer review meeting and rework. Each reviewer finds some defects and makes a report on them during his/her individual check, and then the reviewers participate in the peer review meeting and examine the defects written in the reports. It is said that the reported defects stimulate the other participants to find more defects through their own individual checks. Therefore one of the most important objectives of peer review meetings is to find defects which could not detected in the precedent individual checks.

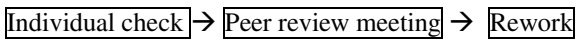

Fig. 1. MAIN PROCESS OF PEER REVIEWS

B. Porter's Experimen result

Porter's experiment collects data from 16 peer reviews for a 24-page software requirements specification document and 16 peer reviews for a 31-page software requirements specification document. An individual check and a peer review meeting respectively spend 120 minutes and 150 minutes on average.

Formula (1) defines a meeting gain rate R_{dm} as the effectiveness of peer review meetings, where N is the total number of detected defects, and N_{dm} is the number of defects which are detected in all the peer reviews meetings.

$$R_{dm} = N_{dm} / N \times 100 \quad (1)$$

Porter's result of R_{dm} is 4.7% for a 24-page document and 3.1% for a 31-page document on average. In addition, he collects data from a different type of peer reviews: two individual checks (without peer review meetings) with the same amount of time as the above experiment's, and he shows the result that the average number of defects without peer review meetings is about twice as large as that with a peer review meeting. Thereby he concludes peer review meetings have no significant contribution to their objective of detecting unfound detects.

C. Our experiment result

Our experiment collects data from six peer reviews for a 12-page software requirements specification document. Each review team has six professional software engineers. An individual check and a peer review meeting respectively spend 30 minutes and 30 minutes on average.

The result of our experiment is shown in Table 1, which remarkably differs from the Porter's of R_{dm}, 4.7% and 3.1%. Our result shows that Porter's conclusion does not always true.

TABLE I. MEASUREMENT RESULTS FOR THE NUMBER OF DEFECT IN PEER REVIEW MEETINGS

Review number	All defects (N)	Defects in peer review meeting (N_{dm})	Meeting gain rate (R_{dm})
1	27	7	25.9
2	13	4	14.8
3	36	10	37.0
4	38	10	37.0
5	29	2	7.4
6	22	3	11.1

We analyze the two experimental results statistically. Table 2 shows F-test results for distribution curve. We get smaller P; 0.00001 than 0.05, and thereby conclude two distributions are statistically different.

TABLE II. RESULTS OF F-TEST

Experiment	Average	Variance	Number of measures	Degrees of freedom	Ratio of Variance	P (right-tail)
Porter's	4.81	10.1	16	15	0.059	0.00001
Our's	22.2	172.2	6	5		

Table 3 shows t-test results. We get smaller P; 0.024 and 0.012 than 0.05, and thereby conclude two averages are statistically different.

TABLE III. RESULTS OF T-TEST

Experiment	Average	t	P (right-tail)	t-value (right-tail)	P (two-tailed)	t-value (two-tailed)
Porter's	4.81	-3.2	0.01	2.02	0.02	2.57
Our's	22.2					

D. Comparison of the two results

We show the differences between our experiment and Porter's experiment and possible explanation on how these differences influenced the findings.

- Human hour for a page

  We use a 12 page software requirement specification. Porter uses both 31 and 24 page software requirements specifications. Porter's document has about twice as many pages as our document does, and the efforts spent for reviewing are also about twice as long as that of our experiment's.

  Porter's experiment: 3 persons, 2 hours, 24 or 31pages

  = 0.25 or 0.19 human hours / page

  Our experiment: 3 persons, 0.5 hours, 12 pages

  = 0.25 human hours / page

  Experimental condition on effort per page was almost the same as ours for 24 pages, and as a result this condition do not influence on the difference.

- Number of participants

  On Porter's experiments, there are three participants although a peer review meeting normally requires four roles, i.e., moderator, reviewer, recorder and reader. It is likely that the three participants in Porter's experiment do not properly perform the four roles. Actually, Porter's peer review meetings do not explicitly mention the existence of trained moderators. On our experiment, six participants attended the peer review and one person acted as a trained moderator. This condition may influence the different conclusion.

- Time of individual check

  In the Porter's experiments, defects detected in two individual checks are twice as many as those in the combination of an individual check and a peer review meeting on average. It can be inferred that the time spent for each individual check in the Porter's experiments is too short to read through documents to be reviewed, which makes subsequent peer review meetings ineffective. Sufficient individual checks make a great influence on increasing the effectiveness of peer review meetings dramatically.

III. CONCLUSION

It had been a contentious subject during the late '90s as to whether peer review meetings make a significant contribution to defect extraction or not. Our findings show that the peer review meeting does indeed make a clear contribution for detecting defects on some cases, which is contradict to the Porter's experiment. The important condition to change peer review meetings into more conducive ones is to spend sufficient time for individual checks. We plan to improve the effectiveness of peer review meetings based on collecting and analyzing further experimental data.

REFERENCES

[1] T. Gilb and D. Graham: Software Inspection, Addison-Wesley, 1993.

[2] P. M. Johnson: Reengineering Inspection, Communications of the ACM, Vol.41, No.2, pp.49-52, 1998.

[3] L. G. Votta: Does Every Inspection Need a Meeting?, in Proceedings of the First ACM SIGSOFT Symposium on Foundations of Software Engineering, pp.107-114, 1993.

[4] A. A. Porter: Assessing Software Review Meetings: Results of Comparative Analysis of Two Experimental Studies, IEEE Transactions on Software Engineering, Vol.23, No.3, pp.129-145, 1997.

[5] A. A. Porter: Comparing Detection Methods for Software Requirements Inspection: A Replicated Experiment, IEEE Transactions on Software Engineering, Vol.21, No.6, pp.563-575, 1995.