Replicated Study of Effectiveness Evaluation of Cutting-Edge Software Engineering*

Yukasa MURAKAMI†, Nonmember and Masateru TSUNODA†a, Member

SUMMARY Although many software engineering studies have been conducted, it is not clear whether they meet the needs of software development practitioners. Some studies evaluated the effectiveness of software engineering research by practitioners, to clarify the research satisfies the needs of the practitioners. We performed replicated study of them, recruiting practitioners who mainly belong to SMEs (small and medium-sized enterprises) to the survey. We asked 16 practitioners to evaluate cutting-edge software engineering studies presented in ICSE 2016. In the survey, we set the viewpoint of the evaluation as the effectiveness for the respondent’s own work. As a result, the ratio of positive answers (i.e., the answers were greater than 2 on a 5-point scale) was 33.3%, and the ratio was lower than past studies. The result was not affected by the number of employees in the respondent’s company, but would be affected by the viewpoint of the evaluation.

key words: replicated study, questionnaire, effectiveness, interest

1. Introduction

Although many software engineering studies have been conducted, it is not clear whether they meet the needs of software development practitioners. Lo et al. [6] pointed out that in many international conferences, very few participants represent the practitioners, raising doubts about whether the research actually meets the needs of practitioners. To eliminate these doubts, previous studies [2], [6] evaluated the effectiveness of software engineering research for practitioners.

We conducted a study that replicates previous research [2], [6]. Shull et al. [8] classified replicated studies into “exact replication” (i.e., the experimental procedures follow the original study) and “conceptual replication” (i.e., research questions are similar, but the experimental procedures differ). This study follows the latter approach, and differs from previous studies in the following ways:

(1) We interviewed practitioners who worked for small and medium-sized companies and analyzed whether the size of their companies affected the effectiveness of the evaluation. In previous studies [2], [6], the majority of respondents were practitioners who worked for large companies such as Microsoft. However, in smaller companies, the number of team members on a project and the scale of the software being developed differ significantly. Therefore, the needs of engineers from smaller companies may also differ.

(2) The respondents evaluated the research, assuming the effectiveness thereof for their own work, instead of for software development in general. Even if one study is considered to be effective in general, it would not necessarily be effective in every case, considering the unique circumstances of each practitioner. This viewpoint makes effective evaluation stricter. In past studies [2], [6], a question asked was “In your opinion, how important are the following pieces of research?” However, the definition of “important” in this context is not clear, and the viewpoint of the evaluation is obscure.

(3) Respondents evaluated each research topic after reading explanatory documents thereon, which are much longer than in past studies [2], [6]. In previous studies [2], [6], the respondents evaluated research by reading a summary that consisted of one or two lines. Conversely, in this study, the respondents evaluated the research by reading an explanatory document consisting of several slides. This is because respondents need understand the research in detail to evaluate the effectiveness thereof for their own work, as described above.

2. Data Collection of Effectiveness Evaluation

Using a questionnaire, the evaluation of the research was collected from practitioners who worked for software development companies. The effectiveness evaluation of the research was assessed via the following procedures: (1) a respondent selected three to five research categories related to their work (at least three categories needed to be selected), (2) the respondent read three documents explaining the research for each selected category, and (3) the respondent evaluated the research, considering “effectiveness for their own work” and “personal interest” on a 5-point scale, 1 being the lowest and 5 the highest.

As previous studies [2], [6] used a 4-point rating scale evaluation, the evaluation was not as generalized [1]. This is because the number of choices is even (i.e., four), and does not allow a neutral rating that is exactly between the highest and lowest. Therefore, we used a 5-point rating scale evaluation in this study.

The category of the research and the papers included in each category were selected based on the International
Conference on Software Engineering (ICSE) held in 2016, which is the premier international conference on software engineering. That is, we used each session of the conference as a research category. The details of these categories are explained in Sect. 3.2. Each category (session) contained approximately four papers. To reduce the task load, respondents were given explanatory documents to read [4] instead of papers. Each document consisted of a few slides that were easy to understand.

The content of the slides might be subjective, depending on its creator, however, it is difficult to avoid the inclusion of subjectivity when research is summarized manually. Past studies [2], [6] also summarized the papers manually in 1-2 lines, but slides used in this study contain much richer information. Therefore, the influence of subjectivity is considered to be relatively small. Although each explanatory document includes a slide titled: “Impressions of the research,” they were not particularly biased to the best of our knowledge.

Each respondent also noted their educational level (e.g., bachelor’s degree), years of work experience, the number of employees in their company (e.g., less than 300; see Table 1), and the developmental processes that they are engaged in. Except for the number of employees, this data was also included in past studies [2], [6].

### 3. Analysis of Effectiveness Evaluation

#### 3.1 Summery of Respondents

There were 16 respondents, each working for a different software development company. Table 1 lists a summary of the respondents stratified by the number of employees in their respective companies. Two respondents had more than three years of experience in the group of employees with less than 300 employees. In the other groups, respondents had one to three years of experience.

To analyze the influence of years of experience, we also list the results omitting respondents with more than three years of experience. On an estimation model constructed based on CoBRA by a research institute [3], the values of a cost factor are decided based on the years of experience, and the threshold is three years. Therefore, we followed this criterion.

Table 1 lists the development process in which respondents were involved, stratified by the number of employees in their respective companies. Most of the respondents were involved in testing and coding, and more than half of the respondents were involved in design. Requirement analysis and development support were not relevant when the number of employees was less than 300, and years of experience were less than three (Table 1, second row). Although the percentages of these categories were low in other cases, they were relevant.

The research categories in the ICSE are listed in Table 2. Due to space limitations, nine categories not related to the respondents’ work were omitted. Each session of the conference was regarded as a research category. There were two testing sessions that were treated as a single session in this study. Each session included approximately four papers.
3.2 Preliminary Analysis

To analyze the effectiveness of research, the research topic should be related to respondents’ work. Therefore, as a preliminary analysis step, the relationship between the work and research categories were analyzed, considering the number of employees (i.e., company size).

Table 2 shows the relationship between the research categories and the respondents’ work. The “Total” column shows the percentage of respondents who confirmed that the research was related to their work. The percentage was calculated without stratification by the number of employees. The other columns were calculated by stratifying the answers by the number of employees. Although multiple choices of topics were allowed, we used the number of respondents and not the total number of answers as the denominator for the percentage because Table 1 also uses the number of respondents as the denominator, and we aim to align the expressions in both tables.

Analysis without stratifying company size: 35% of the categories (9 out of 26 categories) were not relevant to respondents’ work. This is because some research topics are not relevant to ordinal software development. For example, “Compilers and Emerging Trends” and “Energy and Videos” were not relevant. Although some respondents were involved in development support, they did not select categories related to analysis such as “Empirical.” The results suggest that some research categories did not relate to the respondents’ work.

Analysis stratifying company size: “Performance” and “Collaborative” were weakly related to work, when the number of employees of the respondent’s company was less than 300. This also includes answers of respondents whose years of experience was more than three, suggesting that years of experience significantly affected the results. The features of software being developed and the organization of software projects (e.g., the number of engaged developers) differed with the size of the software company, affecting the results.

“APIs” and “Open Source Software” were strongly related to the work of respondents, when the number of employees of the respondents’ company was less than 300, regardless of the respondent’s years of experience. Therefore, the years of experience probably did not affect this result. When the number of employees was not less than 300, the categories were weakly related to the respondents’ work. This is because the scope of software being developed is relatively small for companies with less than 300 employees, and as a result, the influence of APIs and Open Source Software might be relatively high. The results suggest that the relationships between research categories and respondents’ work vary according to the size of the respondents’ company.

3.3 Effectiveness of Research

Analysis focusing on the number of employees: We analyzed the evaluation of the research, focusing on the number of employees in the company. Table 3 shows the percentage of research papers that were rated as category three or higher in terms of effectiveness or interest. In the table, the data is stratified by the size of the company. Previous studies [2], [6] focused on the percentage of respondents who rated the research as three or four using a four-point rating scale. As we used a 5-point rating scale, we focused on respondents who rated the research as neutral or higher.

Regardless of the number of employees, approximately 33% of the respondents rated the effectiveness as three or higher, which is relatively low. Regarding “interest,” more than 60% of the respondents rated it as three or higher. The percentage was slightly higher when the number of employees was less than 300. Although many respondents did not rate the effectiveness of the research highly, they were interested in the research.

Estimation of confidence interval and statistical test of population proportion: The number of respondents was not large; therefore, we estimated the 95% confidence interval of the population proportion of the answers. The sample proportion of answers about interest whose values were three or higher was 64.6%. The lower and upper bounds of the confidence intervals were 49.5% and 77.8%, respectively. Thus, the interest of the respondents was not low, even when we considered the confidence interval.

However, the sample proportion of answers about effectiveness whose values were three or higher was 33.3%. The lower and upper bounds of the confidence interval were 20.4% and 48.4%, respectively. In previous studies [2], [6], 67% and 71% of respondents gave positive evaluations of the research. Even when we compared them with the upper bounds of our results, the proportion was lower than in previous studies.

In previous studies [2], [6], the number of responses and the ratio of positive evaluations were described (9,941 and 67.0% for [2] and 17,913 and 71.0% for [6], respectively). Based on these values, we statistically tested the difference in the proportions of effectiveness between previous studies and our study. The results demonstrated that the ratios of the two studies were statistically different from that of our study at the 5% level of significance ($p = 0.00$). This suggests that if the viewpoint of the evaluation is whether the research is effective for the respondent’s work, the ratio of positive evaluations would be low (especially when

| Number of employees | effectiveness ≥ 3 | interest ≥ 3 |
|---------------------|------------------|--------------|
| Less than 300       | 33.3%            | 72.2%        |
| Less than 300 (Years of exp. < 3) | 33.3% | 58.3% |
| 300 to 999          | 33.3%            | 60.0%        |
| More than 999       | 33.3%            | 60.0%        |
the characteristics of developers were similar to those of the respondents of our study).

**Analysis focusing on academic degree**: In Table 3, the ratio of interest was higher than the ratio of effectiveness. The extent of interest might be affected by the experience of research in graduate schools. Therefore, we stratified the respondents according to their academic qualification (i.e., master’s degree or not). The ratio of positive evaluations of interest was 73.3% when the respondents had a master’s degree, while the ratio was 60.6% when the respondents had a bachelor’s degree or other academic qualification.

That is, the respondents who had a master’s degree were more interested in the research. Thus, the academic qualifications of respondents might have affected their interest in the research. Note that the ratio of positive evaluation of effectiveness was not different among these groups (33.3%).

**Analysis focusing on category**: We analyzed the ratio of positive evaluations, focusing on the research categories. Considering the number of data points, we did not stratify by the number of employees. We selected 13 categories rated by two or more respondents, where the ratio of positive evaluations was 50% or higher. Owing to space limitations, the ratio of each category is omitted.

In terms of effectiveness, only “Performance,” “Synthesis,” and “Collaborative” were highly rated (i.e., the ratio of positive evaluations was 50% or higher). In terms of interest, approximately half (i.e., 7) of the categories were highly rated. The correlation coefficient between the ratio of evaluation of effectiveness and the ratio of interest was 0.47. That is, although effectiveness and interest are related to some extent, the respondents evaluated the research separately considering these two factors. Therefore, even when we focused on each research category, the effectiveness evaluation was not high.

4. Discussion

**Research categorization**: To categorize research papers properly, the category should fulfill the following criteria: “The definition of the category is correct,” “The papers are included in the relevant category,” and “A near equal number of papers is included in each category.” Although we could have defined new categories to classify the papers, it would be difficult to classify the papers into these categories properly, and to balance the number of included papers in each category. Although the classification of papers by the sessions might not be optimal, it appears to have been effective. Therefore, we adopted this classification.

**Years of practical experience**: Previous studies [2], [6] asked whether the aim of the research was relevant, and as a result, each research was predominantly evaluated from this perspective. In our study, the respondents evaluated them from a narrower viewpoint, asking: “Is the research effective for your own work?” Although some respondents had more years of experience, their evaluation of the effectiveness was low. Therefore, it does not seem that years of experience effect the evaluation. Additionally, interest in the research was not low, positing that respondents understood the overview of the research.

Developers with around one year of practical experience are considered to have sufficient knowledge to evaluate the effectiveness of a technology. For example, the estimation model [3] explained in Sect. 3.1, also focuses on whether developers have more than one year of experience to settle values of an independent variable. In addition, a study [5] pointed out that when developers engage in software process improvement activities for one to two years, their activities become more streamlined owing to acquired experience. Therefore, these respondents are considered to be adequate judges of the effectiveness of research, especially when the research relates to the respondents’ work. However, when the respondents had fewer years of experience, they were involved in fewer processes. In future work, respondents with more experience will need to be surveyed to evaluate the effectiveness of a study related to the upstream process.

**Characteristics of companies**: To discuss the analysis results, we focus on the characteristics of companies that the respondents work for. Note that answering the company name was not mandatory; thus, only half of the respondents (i.e., eight respondents) answered their company name. The numbers of employees varied for different companies (i.e., both small and large companies). Additionally, the primary business type among the companies was software development (only one respondent worked as a software developer at a wholesale company).

The clients of the respondents’ companies were mainly domestic. Larger companies tended to develop enterprise software, whereas smaller companies tended to develop embedded or e-commerce software. When the developed software are different, the technologies required for these software may be different. This may affect the preliminary analysis in Sect. 3.2. However, as indicated by Table 3, the evaluated effectiveness of companies of different sizes is the same. Hence, this tendency does not affect the analysis results presented in Sect. 3.3.

Note that we do not have information about companies of the other half of the respondents. Therefore, the prior discussion should only be regarded as a reference.

**Number of respondents**: It is not easy to recruit practitioners as respondents (or subjects) when an experiment takes time to perform. Therefore, in several studies, the number of subjects was the same or smaller than that in our study. For example, in [7], the experiment was performed with 10 subjects. Although the number of data points in the study [7] is different from that in our study, the number of subjects is the same. The number of subjects affects the external validity of the study, and therefore, the extent of the external validity of our study is almost the same as that in previous studies.
5. Conclusion

In this study, practitioners who work for small and medium-sized companies were recruited as respondents to evaluate cutting-edge software engineering research based on the effectiveness for their own work. As a result, the following observations were made:

- When the viewpoint of the evaluation was the effectiveness of the research for the respondent’s own work, the ratio of positive evaluation was low.
- The number of employees in the practitioners’ company did not affect the evaluation of the research. The academic degree of practitioners may affect the evaluation of interest.

Therefore, before an effectiveness evaluation of software engineering research by practitioners is conducted in the future, the viewpoint of the evaluation should be carefully considered.

References

[1] C. Azzara, Questionnaire Design for Business Research, Beyond Linear Thinking–An Interactive Approach, Tate Publishing & Enterprises, 2010.

[2] J. Carver, O. Dieste, N. Kraft, D. Lo, and T. Zimmermann, “How Practitioners Perceive the Relevance of ESEM Research,” Proc. International Symposium on Empirical Software Engineering and Measurement (ESEM), no.56, 10 pages, 2016.

[3] Y. Ishitani, “Software Development Estimation Issues and Solutions,” https://www.ipa.go.jp/files/000004077.pdf, 2012 (in Japanese).

[4] IPSJ/SIGSE Working Group of Activation of International research activities, ICSE2016 study meeting, https://sites.google.com/site/sereadings/icse2016 (in Japanese).

[5] T. Kusanagi, M. Ino, and T. Ishikawa, “Software Process Improvement Activities,” Toshiba Review, vol.61, no.1, pp.6–13, 2006 (in Japanese).

[6] D. Lo, N. Nagappan, and T. Zimmermann, “How practitioners perceive the relevance of software engineering research,” Proc. Joint Meeting on Foundations of Software Engineering (ESEC/FSE), pp.415–425, 2015.

[7] S. Müller and T. Fritz, “Using (bio)metrics to predict code quality online,” Proc. International Conference on Software Engineering (ICSE), pp.452–463, 2016.

[8] F. Shull, J. Carver, S. Vegas, and N. Juristo, “The role of replications in Empirical Software Engineering,” Empirical Software Engineering, vol.13, no.2, pp.211–218, 2008.