Evaluating the BPCRAR Method: A Collaborative Method for Business Process Oriented Requirements Acquisition and Refining

Han Lai[1,2,3], Rong Peng[1,*], and Yuze Ni[1]

[1] State Key Lab. of Software Engineering, Computer School, Wuhan University, Wuhan, China
[2] Chongqing Key Lab. of Electronic Commerce & Supply Chain System
[3] School of Computer Science and Information Engineering
Chongqing Technology and Business University, Chongqing, China
{laihan,rongpeng,niyuze}@whu.edu.cn

Abstract. The goal of requirements elicitation is to understand the stakeholders' needs and constraints, and form the system requirements. But gathering requirements correctly, completely and understandably in a natural way is a great challenge to traditional methods, for requirements analysts always play key roles in the elicitation process dominantly while stakeholders participate in passively. Therefore, strategies that help the identification of requirements based on reducing the requirements analysts' dominance and promoting stakeholders' self-expression and self-improvement are welcomed. This paper reports a controlled experiment to evaluate the Business Process oriented Collaborative Requirements Acquisition and Refining (BPCRAR) method. Compared to JAD, the statistical results show that the requirements elicited by BPCRAR are more complete and understandable. Besides that, the perceived usefulness, ease to learn, and ease of use of BPCRAR are all confirmed by the statistical data got from the questionnaire to the participants.

Keywords: Controlled experiment; Requirements elicitation; BPCRAR; JAD; Evaluation.

1 Introduction

Requirements Elicitation (RE) is a critical process in system/software engineering. Its goal is to understand the stakeholders' needs and constraints, which will be analyzed and specified with requirements [1]. RE should consider the analysis of the organization structure with its business domain and processes. The identification and modeling of organization business processes can (i) help the requirements to represent the real business needs, (ii) reduce the number of redundant requirements, and (iii) be used to guide the development life cycle as a whole [2].

Defining requirements is not a simple knowledge transfer process where requirement engineers elicit and document existing client knowledge [3]. Rather, it is a cognitive

* Corresponding Author.

D. Zowghi and Z. Jin (Eds.): APRES 2014, CCIS 432, pp. 90–104, 2014.

process, in which stakeholders collaboratively find out what has to be done by understanding problems and domains, learning from other stakeholders by negotiating and discussing different viewpoints [4]. The major challenges of defining requirements for software intensive systems are: (i) lack of adequate communication between users and analysts; (ii) users do not have a clear and detailed expectation about their real needs; (iii) each stakeholder has different expectations and describes his/her needs differently. Thus, the requirements identified are always incomplete, ambiguous, and highly volatile. The different communication languages used among stakeholders are prone to misunderstanding.

To meet those challenges, many collaborative RE methods have been proposed. For example, Joint Application Development (JAD), Quality Function Deployment (QFD) and Cooperative Requirements Capture (CRC) have been proposed to reinforce the communication between stakeholders and analysts. But most traditional approaches lack the capabilities to gather requirements clearly and completely in a natural flow [5], as the requirements analysts always dominate the RE process. Therefore, the quality of the elicited requirements heavily depends on the knowledge and the experiences of the requirements analysts.

To reduce the requirements analysts' dominance and promote stakeholders' self-expression and self-improvement, a **B**usiness **P**rocess oriented **C**ollaborative **R**equirements **A**cquisition and **R**efining (BPCRAR) method is proposed [6].

This paper reports the results got from a controlled experiment to evaluate the effectiveness and usefulness of the BPCRAR method. We discuss the quantitative and qualitative findings of our empirical study and their potential for improving the BPCRAR method. These findings indicate that BPCRAR could be considered as a promising method to capture and refine business process oriented requirements.

The remainder of this paper is organized as follows: Section 2 introduces related work about collaborative RE. Section 3 outlines BPCRAR. Section 4 presents the empirical work conducted to evaluate BPCRAR. Section 5 discusses possible threats to validity. Section 6 presents our conclusions and further work.

2 Collaborative Requirements Elicitation

Group work is a common way to elicit requirements collaboratively through the promotion of stakeholders' cooperation and commitment [7]. Brainstorm, JAD and Focus Group are all typical RE Group meeting methods. Group meeting have two modes: face-to-face mode and online mode. Due to the number of stakeholders that may be involved, face-to-face mode group meeting is difficult to organize and schedule; and online mode group meeting becomes popular. The RE methods such as WinWin[8], EasyWinWin [9], CoREA [10] and Athena [11] all support both modes by electronic tools. In addition, many studies have been proposed to facilitate the communication and participation of distributed stakeholders mainly from the computer supported cooperative work, such as RE-specific wikis [12] and iRequire [13]. Last but not least, some studies are focused on utilizing the stakeholders' profile, the social network among stakeholders and data mining technology to generate effective recommendation mechanism to promote the collaboration in RE, such as [14] [15].

All these studies concentrate on facilitating the stakeholders' participation. The BPCRAR method is also a collaborative method and mainly focused on the business process oriented requirements' acquisition and refining. It is a solution about how to refine requirements from group stories to formal expression (i.e. business process models) based on progressive refinement.

3 Overview of BPCRAR

The BPCRAR method is proposed to promote stakeholders' self-expression, self-improvement and collaboration. It adopts the group storytelling [11], dialogue game [16], and narrative network modeling (NNM) [17] to facilitate collaboration and communication. The overview of BPCRAR is shown in Figure 1. It consists of four activities: "group storytelling", "create abstraction", "build formal presentation", and "dialog game". The first three activities form a fountain model [18] from concreteness to abstraction and the activity of "dialog game" is a sub activity throughout the three, which provides the interaction rules and acts as a refining wheel to guide stakeholders' expression. The whole process and each activity in the process can be iterated until the final artifacts are gained and validated.

There are three essential roles in BPCRAR: Teller, Facilitator and Modeler.

- **Tellers** are stakeholders whose expectations are crucial to the success of the system, such as customers, users and domain experts. They can tell their expectations, activities or knowledge about the system to be built by a series of stories.
- **Facilitators** are experienced professionals who mediate the processes of telling stories, link facts, and help to produce the first level of abstraction.
- **Modelers** are assigned to qualified requirement analysts who can develop the graphic business models based on the abstractions extracted from stories.

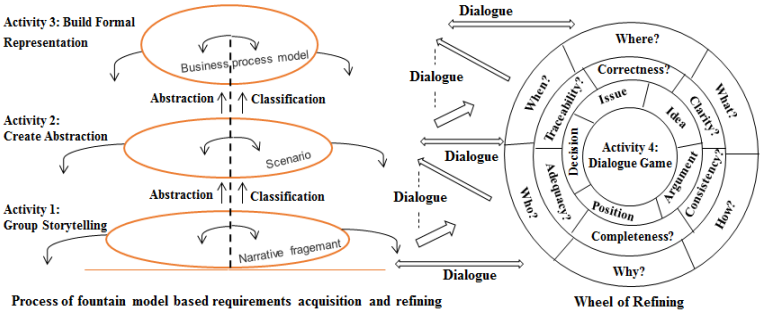

Fig. 1. Overview of BPCRAR Method

The **Activity name**, **Motivation**, **Subjects**, **Objects**, and **Results** of each activity are summarized and based on Activity Theory (AT) [19] in Table 1. The main **Actions** and **Goals** in each activity are further described in Table 2. Due to space limitations, the steps (mainly including **Community**, **Division of labor**, and the **Rules**) taken by BPCRAR are elaborated using the AT in [6].

Table 1. The Motivation, Subjects, Objects and Results of each activity

Activity	Motivation	Subject	Object	Result
A1: Group Storytelling	Acquire stakeholders' needs	T, F, M	Knowledge, expectations and experiences	Stories
A2: Create abstraction	Identify the scenarios Tellers refer to	F, M, T	Stories	BPEs/NNMs
A3: Build formal presentation	Describe requirement formally by business process model	M, F, T	BPEs/ NNMs	BPM
A4: Dialog game	Provide rule for interaction and refine requirement	M, F, T	Issues in stories, BPEs, BPMs and NNMs	Ideas for solving the issues
Note: Teller (T); Modeler(M); Facilitator(F); Business Process Elements (BPE); Business Process Model(BPM)				

Table 2. The main Actions and Goals in each activity

Activity	Action	Goal
A1:Group Storytelling	Set Theme	Define the scope of the RE process
	Storytelling	Capture stakeholders' needs
	Perform dialogue game	Provide triggers for interaction and refining acquired stories
A2: Create abstraction	Read stories and annotate	Set up traceable annotation for abstraction creation in each story
	Extract BPEs based on stories	Create the first level abstraction of business process
	Build NNMs	Capture actual and potential paths of business process
	Perform dialogue game	Provide triggers for interaction and refining business process
A3:Build formal presentation	Build BPMs	Create formal description of requirements
	Perform dialogue game	Provide triggers for interaction and refining business process
	Validate the BPM	Reach consensus
A4: Dialog game	Raise Issues	Introduce new discussion topics to address issues identified
	Propose of Ideas	Generate solutions to resolve the issues
	Discuss proposed Ideas	Clarify the reasons to support or counter proposed Ideas
	Vote	Evaluate the proposal and reach consensus
	Make decision	Select final Idea(s) to resolve the issue

4 The Controlled Experiment

Before the controlled experiment, we have conducted two case studies to test the usability of BPCRAR in June 2013 with CS graduate students and PHD candidates in Wuhan University. The results of the case studies help us to improve BPCRAR. The controlled experiment was designed to validate the feasibility of BPCRAR.

4.1 Experiment Design

The experiment design refers to the guidelines proposed by Wohlin [20]. According to the Goal-Question-Metric (GQM) [20], the goals of the experiment is to (i) compare the completeness and understandability of the requirements artifacts elicited by BPCRAR and JAD to evaluate their effectiveness; and (ii) evaluate the Perceived Usefulness (PU), Perceived Ease of Learning (PEOL) and Perceived Ease of Use (PEOU) of BPCRAR from the viewpoints of Tellers, Facilitators, and Modelers, (iii) identify issues to be improved in BPCRAR. JAD was chosen as the basis for a comparative measure of effectiveness, because it is one of the most well-known industrial collaborative RE method and can be executed by the CS undergraduate students [21].

The experiment includes 3 independent variables and 5 dependent variables, as showed in Table 3. The chosen dependent variables are based on the requirements engineering process evaluation framework [22] and the collaborative process quality evaluation framework [23].

Table 3. Independent variables and dependent variables in the controlled experiment

Variable name	Type	Description
Role	IV	Teller, Facilitator, and Modeler
Method	IV	BPCRAR and JAD
Object	IV	O1, O2 (as in Section 4.2)
PU	DV	The degree to which participants believe that the technology could improve his/her performance at work
PEOL	DV	The degree to which participants believe that learning a particular RE method is effort-free
PEOU	DV	The degree to which participants believe that using a particular RE method is effort-free
DOC	DV	The ratio of all needs are covered by the specified requirements
DOU	DV	The degree to which the requirements are understandable
Note: Independent Variable (IV); Dependent Variable (DV); Degree of Completeness(DOC); Degree of Understandability (DOU)		

According to the experiments' goals, the research questions and hypothesizes are:

RQ1: Is BPCRAR perceived useful from the viewpoints of Tellers, Facilitators, and Modelers? And the following hypotheses are:

$H1_0$: BPCRAR is perceived not useful from the viewpoints of Tellers.

$H2_0$: BPCRAR is perceived not useful from the viewpoints of Facilitators.

$H3_0$: BPCRAR is perceived not useful from the viewpoints of Modelers.

RQ2: Is BPCRAR perceived easy to learn from the viewpoints of Tellers, Facilitators, and Modelers? And the following hypotheses from the viewpoints of Tellers, Facilitators and Modelers can be assigned as $H4_0$, $H5_0$ and $H6_0$ accordingly.

RQ3: Is BPCRAR perceived as ease to use from the viewpoints of Tellers, Facilitators, and Modelers? And the following hypotheses from the viewpoints of Tellers, Facilitators and Modelers can be assigned as $H7_0$, $H8_0$ and $H9_0$ accordingly.

RQ4: Does BPCRAR produce requirements that are more complete than JAD?

$H10_0$: There is no significant difference between the degrees of completeness of requirements captured by BPCRAR and JAD.

RQ5: Does BPCRAR produce requirements that are more understandable than JAD?

$H11_0$: There is no significant difference between the understandability of the requirements captured by BPCRAR and JAD.

To answer the RQ1, RQ2, and RQ3, the questionnaire includes 3 sets of closed-questions (items) as shown in Table 4, 5 and 6. These closed-questions adopt a 7-point Likert scale (1-Strongly Disagree, 4- Neutral, 7- Strongly Agree). The questions to the same independent variable were randomized to prevent systemic response bias. In addition, in order to ensure the balance of questions in the questionnaire, half of the questions were written in negative sentences to avoid monotonous responses [24]. In this experiment, we deal with Likert scale as interval data. According to Wohlin [20], the Mean should be employed as a measure of central tendency and standard deviation. Each subjective dependent variable was quantified by calculating the arithmetical mean of its closed-question values.

Table 4. Questions about the PU of BPCRAR

Code	Question
PU1	The "Group storytelling" activity is helpful to acquire stakeholders' needs
PU2	The "Create abstraction" activity is helpful to identify the scenarios
PU3	The "Build formal presentation" activity is helpful to formal requirement expressions
PU4	The "Dialog game" activity is helpful to provide rules for interaction and refine requirements

Table 5. Questions about the PEOL of BPCRAR

Code	Question
PEOL1	The rules & DoL of "Group storytelling" activity are easy to learn
PEOL2	The rules & DoL of "Create abstraction" activity are easy to learn
PEOL3	The rules & DoL of "Build formal presentation" activity are easy to learn
PEOL4	The rules & division of labor of "Dialog game" activity are easy to learn
Note: Division of Labor (DOL)	

Table 6. Questions about the perceived ease of use of BPCRAR

Code	Question
PEOU1	The "Group storytelling" activity is easy to participate in
PEOU2	The "Create abstraction" activity is easy to participate in
PEOU3	The "Build formal presentation" activity is easy to participate in
PEOU4	The "Dialog game" activity is easy to participate in

For answering the RQ4 and RQ5, the dependent variables, completeness and understandability, are rated by an expert panel respectively. This panel was formed of three Software Engineering professionals with considerable industry experience in reviewing requirements specifications. For reducing the significant discrepancies based on subjective judgment, they first provided an initial list of requirements on the given topic based on their own expertise as a baseline. In the process of reviewing, the baseline was evolved when new requirements in the artifacts proposed by subjects were confirmed by the panel. And the final version was regarded as the baseline to judge the degree of requirements completeness and understandability delivered by each group. Moreover, the DOC and DOU were not only judged by the final artifacts, BPM models, delivered by the Modeler in each group, but also referred all the artifacts delivered in each step. The evaluation was based on three dimensions, syntactic quality, semantic quality and pragmatic quality independently and rated on 5-Likert scale (5-Well above average; 3-Average; and 1-Well below average) [25].

Moreover, the questionnaire has 3 open-questions respectively from PEOL, PEOU, and PU to get feedback from participants.

4.2 Experiment Implementation

The experiment was planned as a balanced within-subject design with a confounding effect, signifying that the same subjects use both methods in a different order and with different experimental objects as shown in Table 7. The method in [26] is adopted to design the experiment. 8 group subjects participated in. Due to the participants' availability, the experiment conducted 2 times. The second experiment (EXP. 2) was strict replications of the first experiment (EXP. 1) with the change of subjects. Strictly replicated experiment also increases the confidence of the experiment validity.

Two experimental objects O1 and O2 were selected. O1 is "Online course enrollment in university" and O2 is "Online train tickets booking". Both objects are familiar to the subjects and the subjects can use BPCRAR and JAD to elicit the requirements of O1 and O2 in 120 minutes.

The experiment was conducted in an Advanced Software Engineering course in November 2013 at the Chongqing Technology and Business University. 60 third-year BSc students major in CS participated in the experiment. The participation was voluntary and the participants were awarded bonus points in their software engineering courses in return. We created 8 groups (7 participants per group), and had 4 alternates in case promised subjects did not show. Finally, the 4 alternates were not used. Each group involved 3 Tellers, 1 Facilitator, 1 Modeler, 1 Scribe, and 1 Observer in both BPCRAR and JAD. The Scribes were responsible for documenting the information in

the sessions. The Observers took charge of supervising the experiment process complied with the instructions. 56 participants were assigned different roles based on their communication capability, prior modeling knowledge, domain knowledge, speed and clarity of writing, and the participation willingness gathered by background questionnaire. They were randomly assigned to each group based on the role assignment. Several documents were designed as instrumentation for the experiment: training slides, method guidelines, data collection tables and questionnaires. The experimental period of each group was 120 minutes. But the experiment can last a little longer if necessary to avoid ceiling effect [27]. After the experiment done, each participant (except the Scribe and Observer) was asked to fill out the questionnaires. To achieve the effect of single-blind experiment, the JAD groups were also asked to do the similar questionnaire, although the data would not be analyzed.

Table 7. Schedule of the controlled experiment

<table>
<tr><td colspan="6">EXP.1</td></tr>
<tr><td></td><td colspan="2">Group A</td><td colspan="2">Group B</td><td>Group C</td><td>Group D</td></tr>
<tr><td rowspan="2">1st Day
(150 min)</td><td colspan="5">1. Introduce Requirements Engineering for all
2. Introduce JAD for all
3. Introduce BPCRAR for all
4. Introduce BPMN for Modeler
5. Train the Scribes to stenograph</td></tr>
<tr><td colspan="5">Questionnaire on PEOL</td></tr>
<tr><td rowspan="3">2ndDay
(20 + 120 min)</td><td colspan="3">Review BPCRAR</td><td colspan="2">Review JAD</td></tr>
<tr><td>BPCRAR to O2</td><td colspan="2">BPCRAR to O1</td><td>JAD to O2</td><td>JAD to O1</td></tr>
<tr><td colspan="5">Questionnaire on PU and PEOU</td></tr>
<tr><td rowspan="3">3rdDay
(20 + 120 min)</td><td colspan="2">Review JAD</td><td colspan="3">Review BPCRAR</td></tr>
<tr><td>JAD to O1</td><td>JAD in O2</td><td>BPCRAR to O1</td><td colspan="2">BPCRAR to O2</td></tr>
<tr><td colspan="5">Questionnaire about PU and PEOU</td></tr>
<tr><td colspan="6">(A week later)</td></tr>
<tr><td colspan="6">EXP. 2 (Strictly replicated EXP. 1)</td></tr>
<tr><td colspan="6">Experts review and rate</td></tr>
<tr><td colspan="6">Note: O1: Online train tickets booking; O2:Online course enrollment at university</td></tr>
</table>

The formal JAD protocol consists of five stages: “Project definition”, “Background research”, “Pre workshop preparation”, “The workshop”, and “Final documentation”, and might execute several days [28]. Due to the time constraints, the subject groups in JAD were required to perform only “The workshop”, and “Final documentation” stages. The materials needed in the first three stages have been directly provided.

Significant differences between the executing processes of BPCRAR and JAD in the experiment are (i) JAD uses the structured brainstorming [28], whereas BPCRAR uses group storytelling to capture users' needs; (ii) JAD is mainly organized and dominated by the Facilitator, while BPCRAR uses dialogue games as the negotiation rules to guide the requirements negotiation; (iii) NNM are introduced in the BPCRAR to capture actual and potential paths of business processes.

4.3 Experiment Results

After the whole two experiments execution, the experts reviewed the requirements artifacts submitted from each experimental group. If a requirement not in the initial list given by the experts, the experts will determine whether it can be a "Realizable" or "false-positives". "Realizable" includes several notions: (i) useful for at least one stakeholder, (ii) technically already implemented or implementable, and (iii) socially and legally implementable. A requirement was considered "false-positive", e.g. beyond the target scope of the requirement elicitation or not "Realizable". Repeat requirements are considered only once. Discrepancies in this review were solved by consensus.

In this section, the experimental results, the effect of the orders of experiment methods and objects, and the Grader inter-rater reliability are all analyzed quantitatively and all the hypotheses are tested by SPSS V19, with significance level $\alpha = 0.05$. In addition, qualitative analysis is applied to analyze the answers to the open questions in the questionnaire.

4.3.1 Quantitative and Qualitative Analyses

After the EXP.1 and EXP.2 were performed, the experts reviewed all the artifacts submitted. After thorough consideration and discussion, a total of 53 and 46 requirements are confirmed by experts on the experiment object O1 and O2 respectively.

Table 8. Overall results about the each perceived measure items

Perceived Measures	Question	Teller (n=24)		Facilitator (n=8)		Modeler (n=8)	
		Mean	STD	Mean	STD	Mean	STD
PU	PU1	5.17	1.34	5.13	0.99	5.25	1.28
	PU2	4.96	1.04	5.38	0.92	5.13	1.25
	PU3	4.96	1.04	5.13	0.84	5.25	0.89
	PU4	5.17	0.96	5.25	0.89	5.13	1.36
PEOL	PEOL1	4.96	0.96	**4.63**	1.06	4.88	0.64
	PEOL2	4.88	0.85	4.75	0.89	4.88	0.84
	PEOL3	4.71	0.86	4.75	0.89	5.00	0.76
	PEOL4	**4.54**	0.88	4.88	0.84	5.00	1.07
PEOU	PEOU1	5.13	0.95	4.88	1.46	4.88	1.25
	PEOU2	5.13	0.80	**4.50**	1.31	5.00	0.93
	PEOU3	4.79	0.98	5.38	0.92	5.25	1.17
	PEOU4	4.79	1.02	4.75	0.89	**4.13**	0.84

In order to enlarge sample size to get hypothesis testing by statistical test method [29], the corresponding data from EXP.1 and EXP.2 are merged.

Table 8 shows the results of descriptive statistics for each perceived items of different roles. The mean scores are all superior to 4 points (neutral score in the 7-point Likert scale), which indicate that Tellers, Facilitators and Molders all showed positive attitude, "slightly Agree", toward the PU, PEOL and PEOU for each activity in BPCRAR. Furthermore, the mean and standard deviation of each perception-based variable from each role is calculated to analyze the whole method respectively as shown in Table 9. The results show that different roles of the subjects showed positive attitude, "slightly agree", toward the PU, PEOL and PEOU for the whole BPCRAR.

The hypotheses H1-H9 were tested by verifying whether the scores that the subjects assign to the PU/PEOL/PEOU are significantly better than the neutral score on the Likert-scale. Shapiro-Wilk test is adopted to test the normality of the data distribution. The data of H1 to H4 and H6 to H9 are normally distributed (p-value ≥ 0.05), therefore, one-tailed one sample t-test are adopted; and the data of H5 are not normally distributed (p-value < 0.05), thus, the Wilcoxon signed-rank test was adopted to test H5. The results in Table 10 state clearly to reject all the hypotheses, namely that the subjects perceived the BPCRAR as easy to learn, easy to use, and useful.

Table 9. Summary of the results of the perceived-based variables

Dependent variable	Teller (n=24)		Facilitator (n=8)		Modeler (n=8)	
	Mean	STD	Mean	STD	Mean	STD
PU	5.06	0.83	5.22	0.73	5.17	1.13
PEOL	4.77	0.75	4.75	0.82	4.94	0.73
PEOU	4.96	0.62	4.88	0.89	4.81	0.86

Table 10. Hypothesis test for perception-based variables

	p-value	Whether to reject null hypothesis
H1	0.000(<0.05)	Yes(BPCRAR is perceived as useful by Tellers)
H2	0.001(<0.05)	Yes(BPCRAR is perceived as useful by Facilitators)
H3	0.011(<0.05)	Yes(BPCRAR is perceived as useful by Modelers)
H4	0.000(<0.05)	Yes(BPCRAR is perceived as ease to learn by Tellers)
H5	0.013(<0.05)[a]	Yes(BPCRAR is perceived as ease to learn by Facilitators)
H6	0.004(<0.05)	Yes(BPCRAR is perceived as ease to learn by Modelers)
H7	0.000(<0.05)	Yes(BPCRAR is perceived as ease to use by Tellers)
H8	0.014(<0.05)	Yes(BPCRAR is perceived as ease to use by Facilitators)
H9	0.016(<0.05)	Yes(BPCRAR is perceived as ease to use by Modelers)
[a] Result obtained with the 1- tailed Wilcoxon signed rank test		

Table 11 shows the descriptive statistics for the performance variables in each experiment. Bold cells indicate that the completeness and understandability of requirements elicited by BPCRAR are higher than those elicited by JAD in both EXP.1 and

EXP.2. As the distribution was normal, the parametric one-tailed t-test was applied to verify the significance of the means. The results in Table 12 state clearly to reject the null hypotheses $H10_0$ and $H11_0$. To guarantee the scorer reliability, the Kendall's coefficient of concordance was adopted to judge the inter-rater reliability of three experts, which validate the reliability (**Completeness**: $W = 0.754$, p =0.003; **Understandability**: $W = 0.718$, p =0.006).

Table 11. Descriptive statistics for the performance variable

	Performance Measures	Method	Min	Max	Mean	STD
EXP.1	Completeness	BPCRAR	2.67	3.67	**3.17**	0.43
		JAD	2.00	3.33	2.58	0.57
EXP.2	Completeness	BPCRAR	3.00	4.00	**3.50**	0.43
		JAD	2.00	3.67	2.83	0.69
EXP.1	Understandability	BPCRAR	2.67	4.00	**3.17**	0.58
		JAD	2.00	3.33	2.50	0.58
EXP.2	Understandability	BPCRAR	2.67	4.00	**3.33**	0.54
		JAD	2.33	3.67	2.92	0.57

Table 12. Hypothesis test for the performance variables

	p-value	Whether to reject null hypothesis
H10	0.016 (<0.05)	Yes (BPCRAR produces requirements more complete than JAD)
H11	0.035 (<0.05)	Yes(BPCRAR produces requirements more understandable than JAD)

To test the effect of the order of both independent variables, RE methods and experimental objects, the method in [26] is adopted.

Suppose difference function $Diff_x = observation_x(A) - observation_x(B)$, where x denotes a particular subject group, and A, B are the two possible values of one independent variable. We created Diff variables from each performance dependent variable. And the statistic results show that the orders of the independent variables have no significant influences on the dependent variables as shown in Table 13 (all the p-values obtained are greater than 0.05).

Table 13. The effect of the orders of methods and experimental objects

Orders	Dependent variables	EXP.1	EXP.2
Methods	Completeness	No(0.937)[a]	No(0.394)
	Understandability	No(0.394)[a]	No(0.699)[a]
Experimental objects	Completeness	No(0.818)[a]	No(0.515)
	Understandability	No(0.515)	No(0.687)
[a] Result obtained with the Mann-Whitney non-parametric test			

Finally, a qualitative analysis was performed to analyze the answers to the open-questions in the questionnaire. Most of the subjects confirmed that BPCRAR promoted the discussions by clarifying the DoL and getting people more involved. However, 2 Facilitators indicated that the "Create abstraction" was heavy workload, and 1 Modeler indicated that it was not easy to remember and follow the instructions in "Dialogue Game". The participants suggested that BPCRAR might be more user-friendly if appropriate tools could be adopted to fulfill the tasks like role assignment, story recording and annotation, and group negotiation. Last but not least, participants suggested that more detailed guidelines and typical examples should be provided to facilitate the execution effectively and efficiently.

4.3.2 Summary of Results

For perception-based measurement, the analysis indicates that all the null hypotheses ($H1_0$-$H9_0$) are rejected. Namely, BPCRAR is perceived useful, easy to learn, and easy to use in requirements acquisition and refining from the viewpoint of tellers, facilitators and modelers respectively. However, the relatively high standard deviations existed in the Table 8 indicate that a few participants have different opinions. With regard to PEOL, the mean of tellers' feedbacks on PEOL4 (Dialogue game) and the mean of facilitators' feedbacks on PEOL1 (Group storytelling) are 4.54 and 4.63 respectively, which are relatively low. This implicates that more detailed guidelines and examples should be provided in these activities, which complied with the feedback got from the answers to the open-questions. With regard to PEOU, the mean of facilitators' feedbacks on PEOU2 (Create abstraction) and the mean of modelers' feedbacks on PEOU4 (Dialogue game) are 4.54 and 4.13 respectively, which is relatively low. It suggests that computer-aided functions should be provided to improve the effectiveness and reduce the workload of these activities.

For performance-based measurement, the results of experiments indicate that all the null hypotheses ($H10_0$-$H11_0$) are rejected. Namely, BPCRAR produces requirements more completely and understandably than JAD. In addition, the results from each experiment (as shown in Table 11) indicate that BPCRAR is superior to JAD in terms of minimum, maximum and average. Furthermore, the means of two performance indicators of BPCRAR are greater than 3 (the neutral score), which indicates that the results are superior to the average level. Meanwhile, the standard deviation of BPCRAR in each experiment is smaller than that of JAD, which indicates that BPCRAR is more stable in terms of DOC and DOU.

The discussion above indicates that BPCRAR could be considered a promising method for collaborative requirements acquisition and refining.

5 Threats to Validity

The main threats to the **internal validity** come from: learning effects, subjects' experiences, information exchange among groups, and understandability of the training documents. The differences of learning effects were alleviated by ensuring that each participant applied both method to different experimental objects, and all the possible order combinations were considered. And the effects of the orders of the methods and the experimental objects were evaluated by statistical tests and the results proved its

validation. Subjects' experiences may influence the execution of the experiments. To alleviate this threat, the pre-questionnaire was introduced to guide the subjects' assignment. Besides, we conducted sufficient training for both methods. To minimize the information exchange among groups, each group had a separate room to perform the task. But EXP.1 and EXP.2 took place on two different weeks. It is difficult to guarantee no information exchange happened. In order to alleviate this situation, at least to some extent, the participants were asked to return all the material at the end of each experiment. Finally, understanding biases of the training material were alleviated by clearing up all the misunderstandings in the experiment session.

The main threats to the **external validity** are: using students as subjects, and the objects' selection criteria. In our study, the students are acceptable as subjects since nobody has previous experience with any method. To balance the abilities of each group, the pre-questionnaire was conducted and used as the evidence of grouping. The experiments objects "online train tickets booking" and "online course enrollment in university" are selected because both of them are familiar to the undergraduates and have similar sizes and complexity. In future, conducting the experiments in industrial should be highlighted.

The main threats to the **construct validity** are: measures that are applied in the quantitative analysis and the reliability of the questionnaire. Measures adopted in the quantitative analysis are those commonly employed in empirical RE experiments [22]. The reliability of the questionnaire is tested by the Cronbach test. Questions related to PU, PEOL, and PEOU obtained a Cronbach's alpha values that are all higher than the acceptable minimum (0.70) [20]. One limitation of our experiment is that lack of the investigation on other factors (e.g., traceability, verification, accuracy) may influence the method adoption in practice. Another limitation is the lack of use of Technology Acceptance Model (TAM) [30]. In our study, the questionnaire items in "perceived of usefulness of BPCRAR" and "perceived of ease of use of BPCRAR" dimensions mainly based on measurement of each activity in BPCRAR and not based on TAM. Since TAM is one of the most widely applied theoretical model to study user acceptance and usage behavior of emerging information technologies, and it has received extensive empirical support through validations and replications, we plan to design questionnaire according to TAM in the future replicated experiment.

The main threat to the **conclusion validity** is the validity of the statistical tests applied. This was alleviated by applying the most common tests that are employed in the empirical software engineering [20]. However, more replications are preferred to confirm these results.

6 Conclusion and Future Work

This paper presents a controlled experiment to evaluate BPCRAR. The completeness and understandability are evaluated in comparison to JAD and the statistic results show that BPCRAR is superior to JAD in both aspects. Meanwhile, the PU, PEOL, and PEOU of BPCRAR are evaluated from the viewpoint of different roles by questionnaire. The statistic results show that stakeholders recognize its usefulness, ease to learn, and ease to use. According to the validity analysis, more experimentation should be performed to confirm the results.

As future work, we plan to replicate the experiment by considering subjects with different levels of experiences in RE (e.g. industrial practitioners) and objects in other business domains. In addition, implementing the collaborative RE tool based on the Mediawiki and its extensions to support the method is our next step.

Acknowledgements. The research was supported by the National Natural Science Foundation of China under Grant No. 61170026, the National High Technology Research and Development Program of China (863 Program) under Grant No. 2013AA12A206, and the Chongqing Key Lab. of Electronic Commerce & Supply Chain System Special Fund (2012ECSC0210).

References

1. Konaté, J., Sahraoui, A.E.K., Kolfschoten, G.L.: Collaborative Requirements Elicitation: A Process-Centred Approach. Group Decision and Negotiation 1-31 (2013)
2. de Oliveira, M., Viana, D., Conte, T., Vieira, S., Marczak, S.: Evaluating the REMO-EKD technique: A technique for the elicitation of software requirements based on EKD organizational models. In: 2013 IEEE Third International Workshop on Empirical Requirements Engineering, pp. 9–16. IEEE (2013)
3. Grady, J.O.: System requirements analysis. Academic Press (2010)
4. Sutcliffe, A.: User-centred requirements engineering. Springer (2002)
5. Boulila, N., Hoffmann, A., Herrmann, A.: Using Storytelling to record requirements: Elements for an effective requirements elicitation approach. In: Proceedings of the 2011 Fourth International Workshop on Multimedia and Enjoyable Requirements Engineering, pp. 9–16 (2011)
6. Lai, H., Peng, R., Ni, Y.: GDS4RE Report. Technical report, `http://www.pengronggroup.org/gds4rereport20140120.pdf`
7. Zowghi, D., Coulin, C.: Requirements elicitation: A survey of techniques, approaches, and tools. In: Engineering and Managing Software Requirements, pp. 19–46. Springer (2005)
8. Boehm, B.W., Ross, R.: Theory-W software project management principles and examples. IEEE Transactions on Software Engineering 15(7), 902–916 (1989)
9. Boehm, B., Gr, U., Nbacher, P., Briggs, R.O.: EasyWinWin: A groupware-supported methodology for requirements negotiation, pp. 720–721 (2001)
10. Geisser, M., Hildenbrand, T.: A method for collaborative requirements elicitation and decision-supported requirements analysis. In: Geisser, M., Hildenbrand, T. (eds.) Advanced Software Engineering: Expanding the Frontiers of Software Technology. IFIP, vol. 219, pp. 108–122. Springer, Boston (2006)
11. Laporti, V., Borges, M.R.S., Braganholo, V.: Athena: A collaborative approach to requirements elicitation. Comput. Ind. 60(6), 367–380 (2009)
12. Peng, R., Lai, H.: DRE-specific Wikis for Distributed Requirements Engineering: A Review. In: Proceedings of the 2012 19th Asia-Pacific Software Engineering Conference, pp. 116–126 (2012)
13. Seyff, N., Graf, F., Maiden, N.: End-user requirements blogging with iRequire. In: 2010 ACM/IEEE 32nd International Conference on Software Engineering, vol. 2, pp. 285–288. IEEE (2010)

14. Castro-Herrera, C., Cleland-Huang, J., Mobasher, B.: Enhancing stakeholder profiles to improve recommendations in online requirements elicitation. In: Proceedings of the 17th IEEE International Conference on Requirements Engineering, pp. 37–46 (2009)
15. Lim, S.L., Quercia, D., Finkelstein, A.: StakeSource: harnessing the power of crowdsourcing and social networks in stakeholder analysis. In: Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering, Cape Town, South Africa, pp. 239–242. ACM (2010)
16. Ravenscroft, A., McAlister, S.: Designing interaction as a dialogue game: Linking social and conceptual dimensions of the learning process. In: Interactions in Online Education: Implications for Theory and Practice, pp. 73–90 (2006)
17. Pentland, B.T., Feldman, M.S.: Narrative networks: Patterns of technology and organization. Organization Science 18(5), 781–795 (2007)
18. Henderson-Sellers, B., Edwards, J.M.: The object-oriented systems life cycle. Commun. Acm. 33(9), 142–159 (1990)
19. Kuutti, K.: Activity theory as a potential framework for human-computer interaction research. In: Context and Consciousness: Activity Theory and Human-Computer Interaction, pp. 17–44 (1996)
20. Wohlin, C., Runeson, P., Hst, M., Ohlsson, M.C., Regnell, B., Wessln, A.: Experimentation in software engineering. Springer (2012)
21. Costain, G., McKenna, B.: Experiencing the Elicitation of User Requirements and Recording Them in Use Case Diagrams through Role-Play. Journal of Information Systems Education 22(4), 367–380 (2011)
22. Emam, K., Madhavji, N.H.: An instrument for measuring the success of the requirements engineering process in information systems development, vol. 1, pp. 201–240. Springer, Netherlands (1996)
23. den Hengst, M., Dean, D.L., Kolfschoten, G., Chakrapani, A.: Assessing the quality of collaborative processes, vol. 1, p. 16b (2006)
24. Hu, P.J., Chau, P.Y., Sheng, O.R.L., Tam, K.Y.: Examining the technology acceptance model using physician acceptance of telemedicine technology. J. Manage. Inform. Syst. 16(2), 91–112 (1999)
25. Gemino, A., Wand, Y.: A framework for empirical evaluation of conceptual modeling techniques. Requirements Engineering 9(4), 248–260 (2004)
26. Briand, L.C., Labiche, Y., Di Penta, M., Yan-Bondoc, H.: An experimental investigation of formality in UML-based development. Ieee. T. Software Eng. 31(10), 833–849 (2005)
27. Dag, I.K., Sjøberg, A.B., Arisholm, E., Dybå, T.: Challenges and recommendations when increasing the realism of controlled software engineering experiments. In: Conradi, R., Wang, A.I. (eds.) ESERNET 2001. LNCS, vol. 2765, pp. 24–38. Springer, Heidelberg (2003)
28. Duggan, E., Thachenkary, C.: Higher Quality Requirements: Supporting Joint Application Development with the Nominal Group Technique. Information Technology and Management 4(4), 391–408 (2003)
29. Dyb Aa, T., Kampenes, V.B., Sj, O., Berg, D.I.K.: A systematic review of statistical power in software engineering experiments. Inform. Software Tech. 48(8), 745–755 (2006)
30. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. Mis. Quart. 319-340 (1989)