

Assessing the Impact of Active Guidance for Defect Detection: A Replicated Experiment

Filippo Lanubile¹, Teresa Mallardo¹, Fabio Calefato¹, Christian Denger², Marcus Ciolkowski³
¹ University of Bari, Dipartimento di Informatica, Bari, Italy
² Fraunhofer IESE, Kaiserslautern, Germany
³ University of Kaiserslautern, Kaiserslautern, Germany

Abstract

Scenario-based reading (SBR) techniques have been proposed as an alternative to checklists to support the inspectors throughout the reading process in the form of operational scenarios. Many studies have been performed to compare these techniques regarding their impact on the inspector performance. However, most of the existing studies have compared generic checklists to a set of specific reading scenarios, thus confounding the effects of two SBR key factors: separation of concerns and active guidance.

In a previous work we have preliminarily conducted a repeated case study at the University of Kaiserslautern to evaluate the impact of active guidance on inspection performance. Specifically, we compared reading scenarios and focused checklists, which were both characterized as being perspective-based. The only difference between the reading techniques was the active guidance provided by the reading scenarios. We now have replicated the initial study with a controlled experiment using as subjects 43 graduate students in computer science at University of Bari. We did not find evidence that active guidance in reading techniques affects the effectiveness or the efficiency of defect detection. However, inspectors showed a better acceptance of focused checklists than reading scenarios.

Keywords: Active Guidance, Quality Assurance, Inspections, Reading Techniques, Scenario-based Reading.

1. Introduction

Software inspection is a structured process for the static verification of software documents, including requirements specifications, design documents as well as source code. From the seminal work of Fagan [10] to its variants [17], inspectors first read a software document (that is, the object of verification), on an individual basis, for the purpose of understanding and

defect detection. Reading techniques for analyzing documents [2] are the key for enhancing the effectiveness and efficiency of inspections [21].

Checklist-based reading (CBR) is the most frequently applied reading technique [7]. Checklists requires inspectors to read the document while answering a list of yes/no questions, based on past knowledge of typical defects [1, 6, 10, 12, 14]. CBR is considered a nonsystematic technique [22] because it does not provide a guideline on how to answer the questions.

Scenario-based reading (SBR) techniques have been proposed to support the inspectors throughout the reading process in the form of operational reading scenarios [2, 3, 22]. A scenario consists of a set of activities aimed to build a model plus a set of questions tied to that model. While building the model and answering the questions, the reader documents the defects he or she detects in the document under inspection. Each reader in the inspection team gets a different and specific scenario in order to minimize the overlapping of discovered defects among team members. This increases the inspection effectiveness after defect collection at the meeting.

Examples of SBR include defect-based reading [22] and perspective-based reading [3]. More recently, use-based reading [9] and usage-based reading [26] were introduced.

Various experiments comparing these techniques against each other have been conducted with the goal to determine which reading technique is better in terms of effectiveness and efficiency of the inspection. However, the results of these experiments do not give a conclusive answer to the research question. Some experiments showed that SBR techniques are more effective and efficient than CBR [3, 15, 18, 22, 23, 26], while other experiments failed to show any significant difference between the techniques [9, 11, 20, 24].

SBR has been designed with the following two key factors:

- *active guidance*: giving guidance on *how* to perform the inspection through actively working with the document;
- *separation of concerns*: restricting the focus of a reviewer to a specific aspect of interest; that is, to guide on *what* to inspect [16].

In past experiments, one generic checklist (with neither separation of concerns nor active guidance) was compared against a number of specific reading scenarios (providing both key factors), containing more detailed and different questions. None of the existing studies has investigated the influence of any of the two key factors in isolation. Thus, we do not know whether active guidance, separation of concerns, or their combined effect lead to the improved performance (if any) of a reading technique.

In a previous work [8] we conducted a repeated case study at the University of Kaiserslautern to evaluate the impact of active guidance on inspection performance, using perspective-based reading (PBR) as representative for SBR. In order to isolate the active guidance factor, checklists and scenarios were designed to be similar to each other with respect to separation of concerns (i.e. we gave the checklists the same focus as the PBR reading scenarios and asked the same questions). Thus, the only decisive difference between (focused) CBR and PBR (both are perspective-based) is that PBR gives the inspectors *active guidance* during the inspection.

We now have replicated the preliminary study by means of a controlled experiment at the University of Bari. With respect to the initial study we got a higher number of subjects involved, we changed the experimental design, and used a partially different instrumentation (the requirements document was different and the inspection was tool assisted). In addition to the research questions we also reused the experimental variables, the checklists and reading scenarios from the earlier study.

The remainder of this paper is organized as follows. Section 2 describes the experiment, including the variables, design, threats to validity, instrumentation, and execution. Section 3 presents the results from data analysis. Section 4 compares these results to the preliminary study and discusses differences between the two replications. Finally, conclusions and future research activities are presented in Section 5.

2. The Experiment

We were interested to further assess the effects of active guidance on defect detection. Thus, our research questions are the following:

- RQ1: Does active guidance improve inspection effectiveness?
 RQ2: Does active guidance improve inspection efficiency?
 RQ3: Do the inspectors perceive the active guidance as a valuable means?

In order to assess the impact of active guidance in isolation, we compared inspections using PBR (i.e., getting active guidance) against inspections using “focused” CBR (i.e., a version of CBR that implements separation of concerns without any active guidance). Then, the above research questions can be rephrased as follows:

- RQ1: Is PBR more effective than focused CBR in finding defects?
 RQ2: Is PBR more efficient than focused CBR in finding defects?
 RQ3: Is PBR better appreciated than focused CBR by the inspectors?

We have investigated these research questions by means of a controlled experiment in a classroom environment. The experiment was conducted as part of a web engineering course at the University of Bari. Participants were 43 CS graduate students attending a web engineering course at the University of Bari. Most of them had experienced requirements inspections with an undergraduate software engineering course, but none had any experience with active guidance reading techniques. Students were encouraged to participate at the experiment by rewarding the additional work in terms of extra points for the final grade.

2.1 Variables

The independent variables are the variables whose values (or levels) determine the different experimental conditions to which a subject may be assigned. We manipulated the following independent variables:

- The reading technique. Subjects and then teams apply PBR with active guidance or focused CBR without active guidance. Except this factor the techniques are similar.
- The perspective. A subject assumes a specific perspective among the following: user, designer, and tester. Each perspective focuses on a different aspect of software quality and this holds both for PBR and focused CBR. As a result, checklists and scenarios of the same perspective share the same focus and (approximately) the same questions.

We measured (directly or indirectly) the following dependent variables:

Effectiveness

- Number of defects found (by the team as well as by individual inspectors).
- Ratio of defects found (by the team as well as individual inspectors): the number of defects found divided by the total number of known defects in the document.
- Number of major defects found (by the team as well as by individual inspectors). A major defect would result, if undetected, in a defect in test or in usage (all other defects are considered minors).
- Ratio of major defects found (by the team as well as individual inspectors): the number of major defects found divided by the total number of known major defects in the document.

Efficiency

- Time (in hours) spent for defect detection (by individual inspectors as well as by the team).
- Number of defects found per hour (by individual inspectors as well as by the team).
- Number of major defects found per hour (by individual inspectors as well as by the team).

Subject's Perception

In order to measure the subject's perception of *usability* of the two reading techniques we asked the subjects to state their degree of agreement, based on a 6-point rating scale, to the following statements:

- The *scenario / checklist* was easy to understand.
- The *scenario steps / checklist questions* were easy to remember.
- The *scenario / checklist* was easy to apply.

To measure and compare the subject's perception of *usefulness* of the two reading technique, the subjects were asked to express their own preferences about using a scenario rather than a checklist for a follow-up inspection. We counted the outcome of this free choice.

2.2 Design

The experiment requires to compare active guidance (PBR) vs. lack of active guidance (focused CBR). Because there was only one requirements document to inspect, we could not reuse the experimental plan from the former study and we had to adopt a different design for the experiment.

In the defect detection stage of the inspection, which is performed by inspectors on an individual basis, the experimental plan corresponds to a 2 x 3 factorial design, where the two factors are:

- 1) Reading Technique (levels: PBR, focused CBR);

- 2) Perspective (levels: User, Tester, Designer).

The reading technique and the perspective variables vary between subjects because none of the subjects is exposed to more than one experimental condition.

Table 1 shows the experimental plan: cells include the numbers of subjects which were randomly assigned to the experimental conditions.

Table 1. Experimental plan

		Perspective		
		User	Designer	Tester
Reading Technique	PBR	7	7	7
	Focused CBR	7	8	7

With this design, it is possible to analyze the influence of the active guidance at team level other than at individual level. In fact, we set up 14 inspection teams of which 7 used PBR (the main treatment) and 7 used focused CBR (the control group). Each inspection team was made up of three inspectors who had assigned different perspectives but the same reading technique, plus an inspection leader (one of the authors) who managed the entire inspection process without taking part in defect discovery. There was a residual participant, performing defect detection with the PBR technique from a Designer perspective, who was not considered part of any team.

From a team level viewpoint, the experimental design can be characterized as a simple single-factor experiment with two levels of the factor (PBR and focused CBR).

2.3 Instrumentation

Instrumentation includes the requirements document and the tool used to assist inspectors.

The document to be inspected represents the requirements specification for developing a door control unit (DCU) for a car. The DCU affects the functions of seat positioning, window movement, exterior mirror adjustment and door locking. The software requirements specification was written in natural language (originally in German and then translated to Italian) and adhered to a use-case style format. It was 31 pages long and contained 21 use-case descriptions.

Inspectors were supported by a web-based tool for distributed software inspections [19]. While performing individually the review task, inspectors recorded defects on a XML-based discovery log. The tool provides support for both checklists and scenarios,

which can be individually assigned to inspectors. Furthermore, the tool adopts a reengineered inspection process [25] that replaces the inspection meeting with two new sequential phases: Collection and Discrimination. They are the result of separating the activities of defect collection (i.e., putting together defects reported by individual reviewers) from defect discrimination (i.e., removing false positives), having removed the goal for team activities of finding further defects. The Collection phase is an individual task and requires either the inspection leader or the author himself. The Discrimination phase is the only phase where inspectors may interact in a meeting but it can be skipped to save time and diminish coordination overhead.

2.4 Training

All subjects were prepared to the experiment with a couple of 3-hour lectures: the former on the support tool and the latter on both focused CBR and PBR scenarios. A requirements document for an eCommerce web application was used to let students make practice. Defects which could be found with the help of the reading techniques were discussed in class, and the tool was used to record findings during a trial inspection.

2.5 Execution

Participants performed the inspections in two university laboratories. They used the tool to display the assigned reading technique and record defects found while reading the document. After individual defect detection, students filled in a debriefing questionnaire and then left the room.

In the next inspection phase (Collection), all the individual findings from the same inspection team were merged by the tool into a unique defect inspection list. Then the inspection leader removed redundant defects, with the help of the tool. Because of time constraints, the discrimination stage was left out.

2.6 Data Collection

The total number of defects in the document was not known in advance. Before the experiment we had derived a first master defect list from a careful review performed by the experimenters. We had also categorized defects as major or minor. Then, when inspectors reported a true defect which was not present on the list of known defects, we added this defect to the list and classified it. We finally ended up with 43

defects in the document of which 16 were major defects.

2.7 Threats to Validity

This section discusses the threats to validity that are relevant for our experiment. To rule out the threats we could not overcome or mitigate, other experiments may use different experimental settings, with other threats to validity of their own. Basili et al. [4] discuss how processes, products, and context models have an impact on experimental designs in the software engineering domain.

Threats to internal validity are rival explanations of the experimental findings that make the cause-effect relationship between independent and dependent variables more difficult to believe. We identified the following threats to internal validity:

Selection. The selection threat refers to natural differences in human performance. In our experiment, we reduced selection effect by randomly assigning subjects to reading techniques and perspectives. Because of the limited timeslot for experimentation in the course, we could not adopt a within-subjects design with repeated measurements of the reading technique factor to minimize the effect of high variation in human performance which might mask differences in the reading technique performance.

Plagiarism. Subjects exchanging information during or between experimentation tasks is a typical risk for experiments in an academic context. We minimized this risk by having only one individual review which was performed in parallel by students. Individual reviews occurred in laboratory with teaching assistants monitoring the task. We told students that the inspection results had no influence on their grade except that they participate in the experiment.

Learning. The learning effect should be symmetric between the values of the independent variables, otherwise it tends to interfere with performance. Although we minimized the learning effect by having multiple training sessions before the experiment, covering both reading techniques from all the perspectives, we cannot exclude that learning was still in progress during the experiment and was fairly balanced between the experimental conditions. While subjects were novices with respect PBR they had a previous experience in a former software engineering course with inspecting requirements documents using generic checklists.

Process conformance. Students may not have followed the checklists and the scenarios as it was intended. We tried to control this threat by not giving

grades tied to the inspector's performance. As the usage of the tool drove participants to apply a specific reading technique, we are confident that inspectors did not switch from the assigned reading technique to another one. In fact, when we asked the students to what degree they had followed the instructions, their answers were not affected by the given reading technique.

Instrumentation. Instrumentation deals with the problem that differences in the results may be caused by differences in experimental material. As we compared only reviews of the same requirements document using the same inspection tool, we have overcome this threat.

Threats to external validity are factors that limit the generalization of the experimental results to the context of interest, here the industrial practice of software inspections. For our experiment, we can identify the following threats to external validity:

Representative subjects. Although our students may not be representative of the whole population of software professionals, recent studies have shown that the difference between students and "real" developers may not be as large as assumed [13].

Representative artifacts. The requirements document inspected in this experiment was provided by Daimler Chrysler. Although it is not a real requirements document Daimler Chrysler ensured that it has the complexity of a requirements document of an electronic control unit in a car. So, it may be considered representative of industrial requirements documents based on a user centered view.

Representative processes. The inspection process in this experiment may not be representative of industrial practice. There are actually many variants of the inspection process in the practice and these include tool-assisted reviews. However, our inspections differ from industrial practice of inspections because individual reviews are not performed on subjects' own desk, with possible interruptions, but in a laboratory setting.

3. Results

Due to the changes of the design of the experiment, the collected data were analyzed differently from the former study [8]. The analysis of this controlled experiment was performed in multiple steps:

- 1) Inspection effectiveness and efficiency were compared at the team level to assess whether there are significant differences.
- 2) The same comparison is repeated but test for differences was conducted at the individual level.

- 3) The techniques are also compared with respect to the perceived usability, on the basis of answers to the debriefing questionnaires. We also compare the techniques with respect to their degree of acceptance as expressed by inspectors' selection for a follow-up inspection.

3.1 Analysis of Team Performance

Having the inspection team as the observational data unit, the experimental design is a simple single-factor experiment with two levels of the factor Reading Technique (we had seven teams for each level). Then, for each of the dependent variables that are representations of effectiveness and efficiency constructs we used a t test for independent-samples to evaluate the differences in means between the two groups.

As shown in Table 2, the analysis failed to reveal any significant difference between the two groups (PBR and focused CBR) at the 0.05 p-level. However, the test of mean differences for the Time might be considered significant at the 0.1 level ($p = 0.086$); that is, teams in the focused CBR group spent more time for defect detection than teams using PBR. The equivalent non-parametric test (Mann-Whitney U test) did not provide different results.

Although the inspectors did not communicate with each other as part of inspection, we did not take into account all possible combinations of inspectors as in [5, 26]. The main reason was our intention to exploit tool support for the entire reengineered inspection process.

3.2 Analysis of Individual Performance

Having the inspector as the observational data unit, results were analyzed using a two-way ANOVA, with two between-groups factors. The first factor is the reading technique with two levels (PBR and focused CBR) while the second factor is the Perspective with three levels: User, Designer and Tester.

For each dependent variable (Ratio of defects found, Ratio of major defects found, Time, Number of defects found per hour, Number of major defects found per hour), we consider the following null hypotheses:

H_{01} : No interaction between Reading Technique and Perspective.

H_{02} : No main effect for Reading Technique.

H_{03} : No main effect for Perspective.

As shown in Table 3, the analysis failed to reveal any significant effects for both the independent variables as well as their interaction.

Table 2. Results at the team level

	Mean Focused CBR	Mean PBR	t-value	df	p	Valid N Focused CBR	Valid N PBR	Std. Dev. Focused CBR	Std. Dev. PBR
Number of defects found	16.85714	14.71429	1.132815	12	0.279419	7	7	3.670993	3.401680
Ratio of defects found	0.39203	0.34219	1.132815	12	0.279419	7	7	0.085372	0.079109
Number of major defects found	8.57143	7.28571	1.065605	12	0.307574	7	7	2.370453	2.138090
Ratio of major defects found	0.53571	0.45536	1.065605	12	0.307574	7	7	0.148153	0.133631
Time (hours)	9.280000	8.375714	1.866631	12	0.086569	7	7	0.801124	1.000514
Number of defects found per hour	1.849979	1.783520	0.232831	12	0.819816	7	7	0.547878	0.519762
Number of major defects found per hour	0.943688	0.883522	0.347501	12	0.734234	7	7	0.337022	0.310254

Table 3. Results at the individual level

ANOVA table for Ratio of defects found					
Effect	Sum of Squares	df	Mean Square	F-Value	p
Reading Technique	0.008302	1	0.008302	1.3376	0.2549
Perspective	0.024338	2	0.012169	1.9608	0.1551
Reading Technique * Perspective	0.000766	2	0.000383	0.0617	0.9402
Error	0.229632	37	0.006206		
ANOVA table for Ratio of major defects found					
Effect	Sum of Squares	df	Mean Square	F-Value	p
Reading Technique	0.006079	1	0.006079	0.33645	0.5654
Perspective	0.007418	2	0.003709	0.20528	0.8153
Reading Technique * Perspective	0.002517	2	0.001259	0.06966	0.9328
Error	0.668527	37	0.018068		
ANOVA table for Time					
Effect	Sum of Squares	df	Mean Square	F-Value	p
Reading Technique	0.5764	1	0.5764	2.396	0.1302
Perspective	0.2731	2	0.1365	0.567	0.5718
Reading Technique * Perspective	0.4623	2	0.2311	0.961	0.3919
Error	8.9019	37	0.2406		
ANOVA table for Number of defects found per hour					
Effect	Sum of Squares	df	Mean Square	F-Value	p
Reading Technique	0.3026	1	0.3026	0.1539	0.6970
Perspective	6.8698	2	3.4349	1.7472	0.1883
Reading Technique * Perspective	0.3931	2	0.1966	0.1000	0.9051
Error	72.7412	37	1.9660		
ANOVA table for Number of major defects found per hour					
Effect	Sum of Squares	df	Mean Square	F-Value	p
Reading Technique	0.01559	1	0.01559	0.02143	0.8844
Perspective	0.21034	2	0.10517	0.14459	0.8659
Reading Technique * Perspective	0.02072	2	0.01036	0.01425	0.9859
Error	26.91191	37	0.72735		

3.3 Analysis of the Subject's Perception

In a debriefing questionnaire that the subjects completed after the experiment, we asked them a set of questions regarding the usability of the reading techniques. Questions were partially based on the questionnaire which was submitted to German students in the former study. Here we present the results of this subjective evaluation.

Table 4 shows the degree of agreement to the statement that the *checklists / reading scenarios* are easy to understand.

There was only one subject who was in disagreement with respect to checklists, while four subjects had some difficulty with understanding scenarios. This makes sense, as a checklist does not contain instructions, and the subjects only have to understand the checklist questions.

Table 4. Answers to "reading technique was easy to understand"

Answers	Results			
	checklist		scenario	
abstruse	0	0.00%	0	0.00%
incomprehensible	0	0.00%	0	0.00%
rather incomprehensible	1	4.76%	4	18.18%
rather comprehensible	7	33.33%	10	45.45%
comprehensible	11	52.39%	2	9.09%
plain	2	9.52%	6	27.28%

Table 5 shows the degree of agreement to the statement that the *checklist questions / reading scenario steps* are easy to remember. Most subjects positively acknowledged the sentence but there were a number of subjects who expressed their disagreement (33% for checklists and 45% for scenarios).

Table 5. Answers to "reading technique instructions were easy to remember"

Answers	Results			
	checklist questions		scenario steps	
completely disagree	1	4.76%	0	0.00%
largely disagree	0	0.00%	2	9.09%
rather disagree	6	28.57%	8	36.36%
rather agree	13	61.91%	7	31.82%
largely agree	1	4.76%	2	9.09%
completely agree	0	0.00%	3	13.64%

Table 6 shows the degree of agreement to the statement that the *checklists / reading scenarios* are easy to apply. While eight subjects who had used the scenarios (36%) did not find the scenarios easy to apply, there was only one subject who was in disagreement with respect to checklists. With respect to the previous questions, the answers about easy of application show a stronger tendency in favor of checklists.

Table 6. Answers to "reading technique was easy to apply"

Answers	Results			
	checklist		scenario	
completely disagree	0	0.00%	0	0.00%
largely disagree	0	0.00%	1	4.55%
rather disagree	1	4.76%	7	31.82%
rather agree	16	76.19%	11	50.00%
largely agree	3	14.29%	3	13.64%
completely agree	1	4.76%	0	0.00%

When asked what percentage of defects they expected to have found in the document (answers are shown in Table 7), subjects who had used checklists appeared to be more confident than those who had used scenarios. This might mean a higher trust into CBR than PBR.

Table 7. Answers to "what percentage of defects do you think you have found?"

Answers	Results			
	checklist		scenario	
0-20 %	1	4.76%	1	4.55%
21-40 %	5	23.81%	8	36.36%
41-60 %	4	19.06%	8	36.36%
61-80 %	9	42.85%	4	18.18%
80-100 %	2	9.52%	1	4.55%

Finally, we measured the user acceptance of the reading techniques by explicitly asking subjects which technique they would had preferred to use in a next inspection. The question was not hypothetical because students had to inspect their own requirements documents as part of their project work in the class. Furthermore the question was not asked just after the experiment, as the other questions, but after one month after its end. In the meantime, our subjects got feedback from us on the results of the inspection and had time to discuss with their peers and think about the choice. We solicited answers by email: 40 subjects over 43 responded by indicating their preferred choice

(perspective was assumed to be the same as in the experiment).

To compare the distributions of answers with respect to the original assignments, a 2 x 2 contingency table is shown (Table 8) for which a Chi-Square test can be run. Chi-Square and V-square (the corrected Chi-Square statistic) p-values are both significant at the 0.05 level. Nine subjects (over twenty) who had assigned a scenario in the first time preferred to use a checklist for the next inspection, while all checklist users retained the assigned reading technique.

Table 8. Contingency table for follow-up choices

	checklist	scenario	Row
What I used (freq.)	20	20	40
Percent of total	25.00%	25.00%	50.00%
What I want to use (freq.)	29	11	40
Percent of total	36.25%	13.75%	50.00%
Column totals	49	31	80
Percent of total	61.25%	38.75%	
Chi-square (df=1)	4.27	p=0.0389	
V-square (df=1)	4.21	p=0.0401	

4. Comparison of Replications

In the previous study, results indicated that PBR was more effective but slightly less efficient than CBR. Moreover, the subjects perceived PBR as easier to use as well as more useful than CBR. However, the results were not significant, as we had only three teams in two runs available.

One important result is that we found hints that active guidance pays off only if the document under inspection exceeds a certain complexity or size: PBR was more effective and more efficient than CBR for the most complex document, while CBR was more effective and efficient for the least complex one.

Table 9 compares the two replications. UKL denotes the original study, conducted at the University of Kaiserslautern, while UniBa denotes the replication at the University of Bari. The two replications have a different experimental design and the document inspected were different. Furthermore, the number of subjects in this replication is approximately 3½ times that of Kaiserslautern.

As in the original study, the results of the Bari replication do not show statistically significant differences ($p < 0.05$) between the two reading techniques, except for usefulness in the Bari replication. However, this time, the data had a tendency in favor of CBR: CBR appeared to be more effective, more efficient, and the subjects perceived it as easier to use as well as more useful than PBR.

Table 9. Comparison of the two replications

* Statistically significant results at the 0.05 level are indicated in bold

		UKL	UniBa
CONTEXT	number of data points	3 teams, 12 subjects	14 teams, 43 subjects
	subjects	German students	Italian students
	number of runs	2 runs	1 run
	experimental design	within subjects	between subjects
	number of documents	3	1
RESULTS	effectiveness	PBR > CBR	CBR > PBR
	efficiency	CBR > PBR	CBR > PBR
	perceived usability	PBR > CBR	CBR > PBR
	perceived usefulness	PBR > CBR	* CBR > PBR

Regarding the size and complexity of the inspected document, the DCU requirements document used in this replication (21 use cases) corresponds to the medium-sized document in the original study. As we found hints in the original experiment that the advantage of PBR seemed to be larger for more complex documents, active guidance may not have been necessary for the DCU document. The overhead required by applying PBR (e.g., specifying test cases, writing down a statechart model, or re-specifying a use case model) would only pay off for significantly large and complex documents. This might explain why subjects found CBR easier to use and more useful.

The more positive subjective perception of CBR in this replication may indicate, in the least, that we should examine and improve the PBR scenarios with respect to usability. This may have affected PBR inspectors who completed the defect detection task earlier than CBR inspectors, although drawing models required by PBR scenarios takes is an extra activity which is not present in checklists. Some subjects said

that they felt rather uncomfortable with scenarios and finished to give up. Maybe, because some subjects did not apply PBR very well, that is the reason why they did not find more defects.

5. Conclusions

Although scenario-based reading techniques have been examined much in the literature, no previous experiment has isolated the driving factors of SBR: separation of concerns and active guidance. We are currently investigating whether active guidance alone affects defect detection performance.

In this paper, we have reported on an experiment conducted at the University of Bari, which replicated a former study conducted at the University of Kaiserslautern in Summer 2003 [8]. Both experiments aimed at isolating the factor of active guidance by comparing perspective-based reading (PBR) with focused checklists (CBR) with respect to inspection effectiveness and efficiency, as well as the perceived usability and usefulness of the reading technique. Thereby, focused checklists were quite similar to the PBR perspectives but did not provide instructions (i.e., active guidance).

Both experiments failed to provide statistically significant results, with the exception of perceived usefulness in the Bari replication where subjects perceived CBR to be more useful than PBR.

However, while the data show a tendency towards PBR in the original study (e.g., PBR was slightly more effective, and perceived as more useful and easier to use), the data in the Bari replication shows a contradictory tendency (e.g., CBR was slightly more effective, and perceived as more useful and easier to use). Although we hesitate to draw strong conclusions from a tendency in the data, the consistency of this tendency across the sub-components of effectiveness and perceived usability and usefulness may indicate that there is an underlying hidden factor that has not been considered so far.

One variable that may explain this difference in the tendency is the complexity of the inspected documents. In the original study, where we used three different documents of different size and complexity, we found hints that the advantage of PBR over CBR was larger with more complex documents. As at least one document used in the original experiment is more complex than the document used in this replication, this variable may be important for explaining the results. However, this is currently a hypothesis that needs to be investigated in future.

All in all, we are still unsure what the decisive factor of a reading technique can be. From the

experiments reported here, the influence of active guidance may be more limited than originally postulated by PBR inventors. However, at this point in time, it is too early to draw conclusions. We need further studies and replications that examine the influence of active guidance, as well as other experiments that isolate different factors (e.g., separation of concerns).

Acknowledgements

This study was partially funded by the German Federal Ministry of Education and Research (BMBF) under the grant VFG0004A (QUASAR). Special thanks to all the students who participated in our study.

References

- [1] A.F. Ackerman, L.S. Buchwald, and F.H. Lewski, "Software Inspections: An Effective Verification Process", *IEEE Software*, Vol. 6, No. 3, May/June 1989, pp. 31-36.
- [2] V.R. Basili "Evolving and Packaging Reading Technologies", *Journal of Systems and Software*, Vol. 38, No. 1, July 1997, pp. 3-12.
- [3] V.R. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Sorumgard, and M. Zelkowitz, "The Empirical Investigation of Perspective-based Reading", *Empirical Software Engineering*, Vol. 1, 1996, pp. 133-164.
- [4] V.R. Basili, F. Shull, F. and Lanubile, "Building Knowledge through Families of Experiments", *IEEE Transactions on Software Engineering*, Vol. 25, No. 4, July/August 1999, pp. 456-473.
- [5] S. Biffi, and M. Halling, "Investigating the Defect Detection Effectiveness and Cost Benefit of Nominal Inspection Teams", *IEEE Transactions on Software Engineering*, Vol. 29, No. 5, May 2003, pp. 385-397.
- [6] Y. Chernak, "A Statistical Approach to the Inspection Checklist Formal Synthesis and Improvement", *IEEE Transactions on Software Engineering*, Vol. 22, No. 12, December 1996, pp. 866-874.
- [7] M. Ciolkowski, O. Laitenberger, and S. Biffi, "Software Reviews: The state of the practice", *IEEE Software*, Vol. 20, No. 6, November/December 2003, pp. 46-51.
- [8] D. Denger, M. Ciolkowski, and F. Lanubile, "Does Active Guidance Improve Software Inspections? A Preliminary Empirical Study", *Proc. of the IASTED International Conference on Software Engineering (SE)*, Innsbruck, Austria, February 2004.

- [9] A. Dunsmore, M. Roper, and M. Wood, "The Development and Evaluation of Three Diverse Techniques for OO Code Inspection", *IEEE Transactions on Software Engineering*, Vol. 29, No. 8, August 2003, pp. 677-686.
- [10] M.E. Fagan, "Design and Code Inspections to Reduce Errors in Program Development", *IBM Systems Journal*, Vol. 15, No. 3, 1976, pp. 182-211.
- [11] P. Fusaro, F. Lanubile, and G. Visaggio, "A Replicated Experiment to Assess Requirements Inspection Techniques", *Empirical Software Engineering*, Vol. 2, No. 1, 1997, pp. 39-57.
- [12] T. Gilb, and D. Graham, *Software Inspection*, Addison-Wesley Publishing Company, 1993.
- [13] M. Höst, B. Regnell, and C. Wohlin, "Using Students as Subjects - A Comparative Study of Students and Professionals in Lead-Time Impact Assessment", *Empirical Software Engineering*, Vol. 5, No. 3, November 2000, pp. 201-214.
- [14] W.S. Humphrey, *Managing the Software Process*, Addison-Wesley Publishing Company, 1989.
- [15] O. Laitenberger, C. Atkinson, M. Schlich, and K. El Eman, "An Experimental Comparison of Reading Techniques for Defect Detection in UML Design Documents", *The Journal of Systems and Software*, Vol. 53, 2000, pp. 183-204.
- [16] O. Laitenberger, "Cost-effective Detection of Software Defects through Perspective-based Inspections", *PhD Thesis in Experimental Software Engineering*, Fraunhofer IRB Verlag, 2000.
- [17] O. Laitenberger, and J.M. DeBaud, "An Encompassing Life Cycle Centric Survey of Software Inspection", *The Journal of Systems and Software*, Vol. 50, 2000, pp. 5-31.
- [18] O. Laitenberger, K. El Eman, and T.G. Harbich, "An Internally Replicated Quasi-Experimental Comparison of Checklist and Perspective-Based Reading of Code Documents", *IEEE Transactions on Software Engineering*, Vol. 27, No. 5, May 2001, pp. 387-421.
- [19] F. Lanubile, T. Mallardo, "Tool Support for Distributed Inspection", *Proc. of the 26th Annual International Computer Software & Applications Conference (COMPSAC 2002)*, Oxford, England, IEEE Computer Society, August 2002, pp. 1071-1076.
- [20] J. Miller, M. Wood, and M. Roper, "Further Experiences with Scenarios and Checklists", *Empirical Software Engineering*, Vol. 3, 1998, pp. 37-64.
- [21] A. Porter, H. Siy, A. Mockus, and L.G. Votta, "Understanding the Sources of Variation in Software Inspections", *ACM Transactions on Software Engineering and Methodology*, Vol. 7, No. 1, January 1998, pp. 41-79.
- [22] A. Porter, L.G. Votta, and V.R. Basili, "Comparing Detection Methods for Software Requirements Inspections: A Replicated Experiment", *IEEE Transactions on Software Engineering*, Vol. 21, No. 6, June 1995, pp. 563-575.
- [23] A. Porter, and L.G. Votta, "Comparing Detection Methods for Software Requirements Specification: A Replication Using Professional Subjects", *Empirical Software Engineering*, Vol. 3, 1998, pp. 355-379.
- [24] K. Sandahl, O. Blomkvist, J. Karlsson, C. Krysander, M. Lindvall, and N. Ohlsson, "An Extended Replication of an Experiment for Assessing Methods for Software Requirements Inspections", *Empirical Software Engineering*, Vol. 3, 1998, pp. 327-354.
- [25] C. Sauer, D.R. Jeffery, L. Land, and P. Yetton, "The effectiveness of software development technical reviews: A behaviourally motivated program of research", *IEEE Transactions on Software Engineering*, Vol. 26, No. 1, January 2000, pp. 1-14.
- [26] T. Thelin, P. Runeson, and C. Wohlin, "An Experimental Comparison of Usage-Based Reading and Checklist-Based Reading", *IEEE Transactions on Software Engineering*, Vol. 29, No. 8, August 2003, pp. 687-704.