Empir Software Eng (2006) 11: 119–142
DOI 10.1007/s10664-006-5967-6

REGULAR PAPER

# Perspective-Based Reading: A Replicated Experiment Focused on Individual Reviewer Effectiveness

**José C. Maldonado · Jeffrey Carver · Forrest Shull · Sandra Fabbri · Emerson Dória · Luciana Martimiano · Manoel Mendonça · Victor Basili**

**Editor:** Murray Wood

J. C. Maldonado
Departamento de Ciências da Computação,
Instituto de Ciências Matemáticas e de Computação (ICMC-USP)
Av. Trabalhador São-Carlense, 400-Centro, Caixa Postal 668
13560 São Carlos, SP Brazil
e-mail: jcmaldon@icmc.sc.usp.br

J. Carver (✉)
Department of Computer Science and Engineering,
Mississippi State University, 300 Butler Hall, Box 9637, Mississippi State, MS 39762, USA
e-mail: carver@cse.msstate.edu

F. Shull
Fraunhofer Center for Experimental Software Engineering, Maryland,
4321 Hartwick Road, Suite 500, College Park, MD 20740, USA
e-mail: fshull@fc-md.umd.edu

S. Fabbri
Departamento de Computação,
Universidade Federal de São Carlos,
Rodovia Washington Luiz, km235, Caixa Postal 676, 13565-905, São Carlos, SP Brazil
e-mail: sfabbri@dc.ufscar.br

E. Dória
Universidade do Oeste Paulista, Faculdade de Informática de Presidente Prudente,
Rua José Bongiovani, 700 - Cidade Universitária,
Bloco H - 1° andar - Campus I, CEP: 19050-680 São Carlos, SP Brazil
e-mail: emerson@icmc.sc.usp.br

L. Martimiano
Instituto de Ciências Matemáticas e de Computação (ICMC-USP)
Av. Trabalhador São-Carlense, 400-Centro, Caixa Postal 668
13560 São Carlos, SP Brazil
e-mail: luciana@icmc.sc.usp.br

M. Mendonça
Departamento de Ciências Exatas–NUPERC
Universidade Salvador, Rua Ponciano de Salvador-BA, 41950-
Oliveira, 126 275 Brazil
e-mail: mgmn@unifacs.br

V. Basili
Fraunhofer Center for Experimental Software Engineering and Department of Computer Science,
University of Maryland, College Park, MD 20742, USA
e-mail: basili@cs.umd.edu

**Abstract** This paper describes a replication conducted to compare the effectiveness of inspectors using Perspective Based Reading (PBR) to the effectiveness of inspectors using a checklist. The goal of this replication was to better understand the complementary aspects of the PBR perspectives. To this end, a brief discussion of the original study is provided as well as a more detailed description of the replication. A detailed statistical analysis is then provided along with analysis of the PBR perspectives.

For the individual PBR perspectives, we saw an interesting dichotomy: In the original study there was little overlap among the sets of defects found by each of the three perspectives, while in the replication two of the three perspectives found similar sets of defects on one of the two documents used in the study. Interestingly this document was the only case where the users of PBR were not more effective than the users of a checklist. This result leads to a new hypothesis that the complementary aspect of the PBR perspectives is the characteristic that provides the benefit over other defect detection techniques.

**Keywords** Software inspections · Experimental replication · Laboratory package · Software reading techniques · Requirements documents

## 1. Introduction

Scenario-based reading techniques are an effective way of improving the efficiency and quality impact resulting from software inspections-which are well documented as a highly effective practice for defect reduction (Shull et al., 2002a). Unlike other inspection methods, scenario-based reading techniques explicitly analyze the stakeholders of the inspected document and provide each inspector with an appropriate and targeted quality focus. The major differences seen when applying reading techniques are: 1) Each reviewer uses a unique quality perspective; 2) The review of the document becomes an active rather than passive undertaking; and 3) Each reviewer is focused on well-articulated aspects of quality that are related to the overall quality goals. Studies have shown the benefits of scenario-based techniques for inspecting natural language requirements documents (Basili et al., 1996), (Ciolkowski et al., 1997), (Shull, 1998), formal requirements documents (Porter and Votta, 1998), user interfaces (Zhang et al., 1999), object oriented designs (Laitenberger et al., 2000), (Conradi et al., 2003) and code (Laitenberger et al., 2001). These individual reading techniques are used by inspectors prior to a team meeting (or virtual team meeting). The studies discussed in this paper focus only on the individual defect detection phase. Any results reported for teams of inspectors were gathered from simulated teams, that is defects reported by individuals were aggregated into a team list with no actual team meeting taking place.

Some research has now been done to understand why scenario-based techniques provide the observed benefits. There has been some controversy, for example, over whether the unique quality focus assigned to each reviewer results in less overlap in the defects caught by the reviewers, and hence increases the overall team detection rates. In an early study of Perspective-Based Reading (PBR), a specific reading technique tailored for inspections of natural-language requirements documents, researchers presented subjective results that seemed to indicate that when PBR was being followed with at least minimal process conformance, reviewers who used different perspectives found few defects in common (Basili et al., 1996). However, a later replication of that study used a more rigorous statistical analysis and concluded that there was in fact no significant difference among the different perspectives in terms of the defects that they find (Regnell et al., 2000).

To investigate this discrepancy, we undertook another replication of the study in a different environment. Furthermore, we reanalyzed the data from the original study using a more rigorous statistical approach similar to that used by Regnell et al. Our primary research goal was

> To understand whether the PBR perspectives differ from each other in terms of effectiveness and specific defects found.

Because other studies have shown that a reviewer's effectiveness when using PBR can vary based on the level of his or her experience (Shull, 1998), we also decided to investigate reviewer experience as a possible confounding variable, leading to an additional research goal:

> To understand the impact of a reviewer's experience on his/her effectiveness when using PBR.

### 1.1. Object of Study: PBR

Perspective Based Reading (PBR) is a family of reading techniques designed for inspecting natural language requirements documents developed by the Experimental Software Engineering Group at the University of Maryland (Basili et al., 1996). PBR provides guidance to help an inspector assume the perspective of one of the major stakeholders of the requirements document. The "basic set" of perspectives was defined as: software designer (D), tester (T), and end user (U). The inspector creates an abstraction of the requirements that is relevant to their assigned perspective. For example, a designer creates a preliminary high-level design, a tester creates a set of test cases and a user creates a set of use cases. While creating the abstraction, the inspector uses a series of questions to help them uncover defects. The questions are generated based on a common defect classification. Each question guides the inspector to examine the requirements based on the insight they have gained during the creation of their abstraction.

The set of defect types is not static and can be tailored as necessary for each environment. In current application, it is expected that someone knowledgeable in inspections and/or PBR will perform this tailoring based on expertise and experience. (However, ongoing work is aimed at making this more automated and repeatable.) The instantiation of PBR used in this experiment is reused from a version that was tailored for NASA professionals (Basili et al., 1996).

### 1.2. Research Context

This study was conducted as part of the Readers' Project, a collaboration funded by the US (NSF) and Brazilian (CNPq) national science agencies, with the goal of effectively replicating software engineering experiments (Shull et al., 2002b). A team of researchers from Brazil conducted the replication described here. These researchers were independent from the team of researchers at the University of Maryland who conducted the original study. During the replication, there was interaction between the replicators and the original experimenters. At the conclusion of the study, the replicators from Brazil worked with the original experimenters to analyze and synthesize the data and the results. These two groups of researchers are often differentiated in the paper because of the different tasks performed.

## 2. Short Description of the Original Experiment

The original experiment, conducted at the University of Maryland, compared the effectiveness of teams of subjects using PBR to the effectiveness of teams of subjects using their normal technique for detecting defects in a requirements document. We chose to replicate this study because it was a well designed study, where the treatments allowed multiple variables to be studied, which provided some solid evidence that PBR was effective for inspection teams. We were able to refine the experimental design and goals to investigate other related variables. In addition, the original study left some open questions about PBR that were of interest. The remainder of this section summarizes the original study (Basili et al., 1996).

The research questions (denoted O1–O3, for Original study) were:

O1) If teams of individuals (such as during an inspection meeting) were given unique PBR perspectives, would a larger collection of defects be detected than if each read the document in a similar way?

O2) If individuals read a document using PBR, would they find a different number of defects than if they read the document using their normal technique?

O3) Does a reviewer's experience in his or her perspective affect his or her effectiveness with PBR?

*Subjects*

The subjects were professional software developers from NASA's Goddard Space Flight Center. The study was a within-subjects comparison, in which subjects first applied their usual approach to review two requirements documents, were trained in PBR, and then applied PBR to review two different documents.

*Materials*

Four documents were inspected, two NASA specific documents (NASA A and NASA B) and two generic documents (Parking Garage—PG and Automated Teller Machine—ATM). The documents within each set (NASA or generic) were comparable to each other in length and number of seeded defects.

The techniques used for the inspection were the *Usual Technique* and a *PBR Technique*. The *Usual Technique* was the normal method used by NASA professionals when reviewing requirements documents. This technique was less procedural than PBR in that it was a checklist technique that provided a high-level description of the types of defects for which reviewers should look, but no guidance on how to look for those defects. The checklist items had evolved over time based upon recognizing recurring issues that required clarification from the document authors. There was no procedure or scenario to follow nor any indications about which types of defects might affect which parts of the document.

*Procedure*

As illustrated in Fig. 1, on the first day the subjects received some introductory explanation and then were asked to inspect a NASA artifact and then a generic artifact with their usual technique. On the second day, the subjects were trained in PBR and the perspective to which they had been assigned. They applied PBR first to inspect a generic document and afterwards a NASA document. The order of the documents was assigned randomly to subjects, so that the subjects in Group 1 performed the usual review on NASA A and ATM and the PBR review on NASA B and PG, while subjects in Group 2 did the opposite ordering.

The effectiveness of each review was measured as the percentage of the seeded defects uncovered during the review. Thus, for each subject, the effectiveness during each of the two reviews could be compared to discover whether there was any net improvement due to PBR.

*Results*

Based on the three research questions, the results were as follows:

O1) Teams of subjects using PBR found more defects overall than teams of subjects using the normal NASA technique. This result was statistically significant.

O2) Individual subjects inspecting the generic documents (ATM & PG) found more defects when using PBR than when using their normal NASA technique. This result was also statistically significant. When inspecting the NASA specific documents, individual subjects using PBR found slightly more defects than individual subjects using the normal NASA technique. This result was not statistically significant.

O3) There was no correlation between an inspector's experience in their PBR perspective and their inspection effectiveness.

A laboratory package describes an experiment in specific terms, provides materials for replication, highlights opportunities for variation, and builds a context for combining results of different types of experimental treatments. Laboratory packages build an experimental infrastructure for supporting future replications. They establish a basis for confirming or denying original results, complementing the original experiment, and tailoring the object of study to specific experimental

| | Group 1 | | | Group 2 | | | |
|---|---|---|---|---|---|---|---|
| | Designer | Tester | User | Designer | Tester | User | |
| Usual Technique | NASA A document inspection | | | NASA B document inspection | | | First Day |
| | ATM document inspection | | | PG document inspection | | | |
| PBR Technique | Training in PBR | | | | | | Second Day |
| | PG document inspection | | | ATM document inspection | | | |
| | NASA B document inspection | | | NASA A document inspection | | | |

**Fig. 1** Design of the original study

contexts. A laboratory package is available for this experiment http://www.cs.umd.edu/projects/SoftEng/ESEG/manual/pbr_package/manual.html).

## 3. The Replication

### 3.1. Goals of the Study

As discussed in Section 1, our major research goal was to study whether improved effectiveness due to PBR was a result of different perspectives leading to the detection of different defects. This goal was refined into specific research questions R1 and R2 (where the 'R' indicates that the research question was added for the replication):

R1) Do each of the three PBR perspectives have the same effectiveness and efficiency?
R2) Do each of the three PBR perspectives tend to find the same set of defects?

Before investigating the cause of the improvement due to PBR, we needed to evaluate whether there was in fact any improvement due to PBR in this environment. Therefore we kept research questions O1 and O2 from the original study. Also, having identified a potential confounding factor in reviewer experience (also described in Section 1), we kept research question O3 from the original study. Regarding these questions we introduce only one change: we compared PBR to checklist inspections, the current industry standard, rather than to the technique normally applied by NASA subjects (our subjects did not know the NASA technique).

O1′) Do PBR teams detect more defects than Checklist teams?
O2′) Do individual PBR or Checklist reviewers find more defects?
O3′) Does the reviewer's experience affect his or her effectiveness?

The secondary analysis in the original study showed that PBR was beneficial not only for teams, but also for individuals. So, in the replication we wanted to get further insight into the differences between individual PBR users and individual checklist users. Question R3 was added for this reason.

R3) Do individual reviewers using PBR and Checklist find different defects?

### 3.2. Subjects

Subjects (18 undergraduate students with slightly more than one year of classroom experience on average, from the Software Engineering course at University of São Paulo at São Carlos) were randomly divided into two groups of nine (based on the experimental design in Section 3.4).

The subjects all rated their English-language skills to be at least "medium." The researchers agreed that the level of English was adequate for participation in this study. The subjects in the two groups were not significantly different from each other in terms of experience as software engineers. About half of the subjects had two years or more as developers with only two of them having more than three years. Thus, in general, the subjects' industrial experience was low. Furthermore, the subjects generally had low experience in their assigned PBR perspective.

### 3.3. Materials

Two of the requirements documents, ATM and PG, from the original experiment were used in the replication. The NASA documents were not used because the subjects in the replication were not familiar enough with the domain. The ATM document was 17 pages long with 39 requirements (26 functional and 13 non-functional) and 30 seeded defects. The PG document was 17 pages long with 37 requirements (21 functional and 16 non-functional) and 28 seeded defects. Each subject used a checklist and one of the three PBR perspectives. The checklist was more systematic than the ad hoc procedure used in the original study. In both the pilot study (described below) and the replication, the subjects found new defects that were not found in the original study. The results and analysis presented in this paper include these new defects: ATM with 37 (7 new defects) and PG with 32 (4 new defects).

### 3.4. Procedure

The procedure includes the experimental design, the pilot study, the main steps conducted during the replication and the data collection.

*Design*

The design of the original study was duplicated for the replication with two changes (see Fig. 2). First, the NASA artifacts were not used in the replication. Second, on the first day of the original study, the subjects used their normal technique, while in the replication they subjects used a defect-based checklist.

*Pilot Study*

The replicators in Brazil had not taken part in a replication as part of the original experimenters' group and the level of effort involved in replicating an experiment is high, so a pilot study was done first. The replicators wanted to test the material and concepts in their own environment before running a full experiment. The pilot study was used to better understand the experimental process, including timing, tasks to be executed, documents to be delivered to the subjects and tacit knowledge. These steps were taken to help ensure the quality and conformance of the experimental process.

For example, the replicators did not understand why the same amount of training time was specified for both the usual technique and the PBR approach, since they

| | Group 1 | | | Group 2 | | | |
|---|---|---|---|---|---|---|---|
| | *Designer*<br>**3 Subjects** | *Tester*<br>**3 Subjects** | *User*<br>**3 Subjects** | *Designer*<br>**3 Subjects** | *Tester*<br>**3 Subjects** | *User*<br>**3 Subjects** | |
| Checklist | Training in Checklist | | | | | | First Day |
| | ATM document inspection | | | PG document inspection | | | |
| PBR Technique | Training in PBR | | | | | | Second Day |
| | PG document inspection | | | ATM document inspection | | | |

**Fig. 2** Experimental design

provide different levels of detail and require different levels of background knowledge. For each perspective of PBR, the procedure, questions and defect taxonomy have to be taught. The replicators adjusted the training time but kept it equal for both techniques. It is important to point out that the training is a threat to validity, because the two techniques had varying complexity. This point should be addressed in further experiments.

The pilot study used the same experimental design as the full experiment. Six Masters and PhD students from the Software Engineering Laboratory at University of São Paulo and São Carlos were divided into two groups, with each PBR perspective being used by one subject. These subjects had an average of 5 years of classroom experience as developers.

The data collected in the pilot study was in line with the previous results for the ATM document and inconclusive (i.e., there was no significant difference between the techniques) for the PG document. The quantitative results provided the replicators with confidence that they had achieved a minimal level of understanding to continue with a full study. Aside from identifying missing knowledge, such as the amount of time spent in training, the pilot study gave the replicators an opportunity to better understand how to conduct the study.

Based on our experience, we recommend including a pilot study in any replication process. The pilot study is a useful tool for mastering the underlying concepts, that is the details of the techniques under study, and tacit knowledge, that is information about how to run an experiment that is not recorded in the documentation, and assessing process conformance, to ensure that the study done by the replicators will be comparable the original study, before any significant replication effort is undertaken.

At the conclusion of the pilot study, the main study began. It consisted of the following steps:

1. Subjects filling out the consent form and the background survey;
2. Experimenters training the subjects in the techniques;
3. Subjects applying the techniques;
4. Experimenters collecting data and analyzing it;
5. Experimenters' feedback to the subjects.

*Training*

The training was done in two 2-hour sessions using another artifact, ABC Video Store. The sessions consisted of 30 minutes of theoretical presentation and 90 minutes of practice with the techniques. At the end of the training, the researchers gave the subjects feedback on their performance and the full list of defects for the ABC Video Store document. This feedback allowed the subjects to see the types of defects that they did not uncover and use this information in future applications of the technique.

*Applying the Techniques*

After the training, the subjects applied the Checklist and PBR techniques in sessions of 1 hour and 45 minutes each. On the first day, after receiving training in the Checklist method, the subjects from Group 1 reviewed the ATM document and subjects from Group 2 reviewed the PG document. Each subject was assigned to one of three subgroups for PBR (one for each perspective). On the second day, after receiving training in the assigned PBR perspective, the subjects reviewed the other requirements document. Thus, for each subject, the effectiveness during each of the two reviews could be compared to discover whether there was any net improvement due to PBR.

The subjects performed the inspections in a classroom while the experimenters were present. This setup was slightly different from the original experiment where the subjects were allowed to return to their offices to work. During the inspection of the documents the subjects recorded any defects they found along with a classification for the defect. The subjects did not register the time at which each defect was found. In future studies, we will ask for this information so that we can conduct further analysis on the learning curve of the techniques. Most of the subjects finished the reading activity before the allocated time had elapsed. (For the ATM document 5 of the 18 subjects used the entire 105 minutes. For the PG document only 2 used the entire 105 minutes.)

*Data Collection*

Four metrics were collected during the study. To better address research questions R1–R3, we introduce a new metric, Occurrences of Defects, which measures the uniformity of results among reviewers using the same technique or perspective.

- Defects Found: The number of unique defects found by one or more subjects (i.e., each defect is counted only one time regardless of how many subjects find the defect);
- Occurrences of a Defect: This metric represents the number of times the defect is found, (assuming each subject has the chance to find the defect). The maximum number of occurrences for a defect is the number of inspectors in a group. In each analysis below, groups are defined differently. In Section 3.5.2 there is a group for Checklist inspectors and a group for PBR inspectors. In Section 3.5.6 there is a group for each of the three perspectives. The total number of occurrences of all defects (TotalOc) is calculated as:

  $$\text{TotalOc} = \sum_{i=1}^{n} (x_i)$$

  where $x_i$ is the number of defects found by subject $i$.

  Because occurrences measures the amount of overlap among a group of subjects, for each individual subject, the number of defects and occurrences is the same.
- Effectiveness: The average percentage of defects found by a group of subjects. It is calculated as:

  $$\left(\sum_{i=1}^{n} (x_i/y)\right) * 100/n$$

  where $x_i$ is the number of defects found by the subject $i$, $y$ is the total number of defects in the document and $n$ is the number of subjects in the group.
- Efficiency: The average defects found by each subject per hour. It is calculated as:

  $$\left(\sum_{i=1}^{n} (x_i/k_i)\right) \Big/ n$$

where $x_i$ is the number of defects found by the subject $i$, $k_i$ is the effort (in hours) used by subject $i$ and $n$ is the number of subjects in the group.

## 3.5. Results

Our main interest in this study was to further investigate whether the PBR perspectives were different from each other. Before looking at the results for the individual perspectives, we first wanted to understand the relative effectiveness and efficiency of PBR vs. Checklist. So, we first briefly report the results relative to O1′–O3′ and R3. After that, more detailed results and discussion are provided for research questions R1 and R2.

### *3.5.1. O1′: Do PBR Teams Detect More Defects than Checklist Teams?*

To investigate this question, we posed hypotheses similar to those in the original study. Also, as in the original study, analysis relies on a permutation test, in which team success rates are reasoned about based on computing the overlap of sets of individual reviewers who did not actually work together in a team. The underlying assumption behind such a test is that if PBR and non-PBR reviewers perform differently then teams composed of subjects who all applied PBR (or who all applied a non-PBR technique) should produce different results than teams in which PBR and non-PBR reviewers were mixed.

H0: There is no difference in the defect detection rates of teams applying PBR compared to teams applying the Checklist technique. That is, every successive dilution of a PBR team with a non-PBR reviewer has only random effects on team scores.

Ha: The defect detection rates of teams applying PBR are higher compared to teams using the Checklist technique. That is, every successive dilution of a PBR team with a non-PBR reviewer decreases the effectiveness of the team.

Doing a permutation test, as done in the original experiment, there were 48620 distinct ways to assign the reviewers into groups of 9. The group with no dilution (all PBR reviewers) had the 24769th highest test statistic, corresponding to a p-value of 0.51. Therefore, unlike the original study, we cannot reject the hypothesis H0.

### *3.5.2. O2′: Do Individual PBR or Checklist Reviewers Find More Defects?*

When analyzing the data for the individual inspectors, we first performed a statistical analysis and then a qualitative analysis. The goal of the statistical analysis was to determine whether individual reviewers performed differently when using PBR than when using Checklist. The dependent variables were individual effectiveness and efficiency. Because the experimental groups had the same number of subjects the ANOVA for balanced design was used. This analysis involved two different factors, or treatments: the reading technique (RT: PBR or Checklist) and the requirements documents (DOC: PG or ATM). Three sets of hypotheses were tested with relation to effectiveness and to efficiency.

*Group effect (RT X DOC interaction)*

H0: There is no difference between Group 1 and Group 2 with respect to individual effectiveness/efficiency.

Ha: There is a difference between Group 1 and Group 2 with respect to individual effectiveness/efficiency.

*Main effect RT*

H0: There is no difference between subjects using PBR and subjects using Checklist with respect to individual effectiveness/efficiency.

Ha: There is a difference between subjects using PBR and subjects using Checklist with respect to individual effectiveness/efficiency.

*Main effect DOC*

H0: There is no difference between subjects reading ATM and subjects reading PG with respect to individual effectiveness/efficiency.

Ha: There is a difference between subjects reading ATM and subjects reading PG with respect to individual effectiveness/efficiency.

The results of the ANOVA (shown in Fig. 3) for effectiveness do not allow H0 to be rejected for the group effect (p = .275) or for the main effect RT (p = .354). Therefore, we cannot show that the interaction between the document and the technique or the technique alone have a significant influence on the results. Conversely, H0 can be rejected for the main effect DOC (p = .005), meaning that the document did significantly influence the results. Furthermore, the results of the statistical analysis for efficiency do not allow H0 to be rejected for the group effect (p = .411), the main effect RT (p = .094) or the main effect DOC (p = .121).

To further study the question of whether Checklist or PBR is more effective and efficient, Table 1 summarizes the data collected concerning defects found (union of defects found by individual inspectors) and defect occurrences, as well as average subject effectiveness and efficiency.

In the original study, individuals using PBR were significantly more effective for both PG and ATM (efficiency was not measured). The results of the replication did not show any significant support for those results. For the ATM document, the

**ANOVA Table for Rate (% Defects Found)**

| | DF | Sum of Squares | Mean Square | F-Value | p-value |
|---|---|---|---|---|---|
| RT | 1 | .003 | .003 | .885 | .354 |
| DOC | 1 | .034 | .034 | 8.979 | .005 |
| RT X DOC | 1 | .005 | .005 | 1.233 | .275 |

**ANOVA Table for Efficiency (Defects/hour)**

| | DF | Sum of Squares | Mean Square | F-Value | p-value |
|---|---|---|---|---|---|
| RT | 1 | 10.644 | 10.644 | 2.983 | .094 |
| DOC | 1 | 9.062 | 9.062 | 2.54 | .121 |
| RT X DOC | 1 | 2.477 | 2.477 | .694 | .411 |

**Fig. 3** ANOVA for rate and efficiency (by technique)

**Table 1** Summary of results

| Document | ATM | | PG | |
|---|---|---|---|---|
| Technique | Checklist | PBR | Checklist | PBR |
| Defects Found/Total defects | 15/37 (40.5%) | 21/37 (56.8%) | 20/32 (60.5%) | 14/32 (43.75%) |
| Occurrences of Defects/Total occurrences | 24/333 | 38/333 | 45/288 | 44/288 |
| Effectiveness (% defects found) | 7.21 | 11.41 | 15.63 | 15.28 |
| Efficiency (defects/hour) | 2.00 | 3.62 | 3.53 | 4.10 |

subjects using PBR found a higher percentage of defects than the subjects using the checklist. This result was not statistically significant at the .05 level (p = .143). For the PG document, the subjects using the checklist found a higher percentage of defects on average than the subjects using PBR. In this case, the results was also not statistically significant at the .05 level (p = .911). In terms of efficiency (defects/hour), subjects using PBR were more efficient for both documents. This result was also not statistically significant at the .05 level (ATM p-value = .107, PG p-value = .51). There are two potential explanations for the disagreement in the results:

1. The original study compared PBR to an expert's normal inspection method while the replication compared PBR to a checklist. It is possible that the relationship of PBR to an expert's technique is different than its relationship to a checklist.
2. The cultural differences between the subjects, including informal knowledge about the application domains.

### *3.5.3. O3′: Does the Reviewer's Experience Affect His or Her Effectiveness?*

Similar to the original study, we used a questionnaire to measure the subject's experience in their assigned perspective. The subjects were asked to indicate the number of years experience in specific tasks related to the three PBR perspectives. The relationship between experience and effectiveness was weak (Spearman and Pearson correlation tests were less than .14, far from a high degree of correlation). The range of experience in the subjects was relatively small, i.e., none of the subjects were very experienced. In this case, reviewers with more experience did not perform better than reviewers with less experience.

### *3.5.4. R3: Do Individual Reviewers Using PBR and Checklist Find Different Defects?*

In addition to understanding which technique found more defects, we also addressed the question of whether the users of Checklist and PBR found different defects. Figure 4 provides an overview of the number of unique defects identified by users of each technique. For the ATM document, the two techniques appear to be complementary in that users of each technique found defects that were not found by the other technique. Conversely, for the PG document, the techniques do not

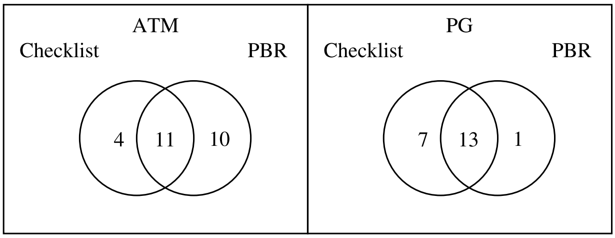

**Fig. 4** Defects found per techniques

appear to be complementary, because the PBR users only found 1 defect not found by the checklist users. Figure 5 shows the data from Fig. 4 in more detail. It shows the number of subjects from each technique type that found each defect (defects that were found by no inspectors are excluded). Overall, the subjects (considering both techniques together) found only 25 out of the 37 ATM defects and 21 out of the 32 PG defects. Therefore, it may be necessary to complement the Checklist and PBR techniques with one or more other techniques to achieve 100% coverage of the defects.

We also investigated whether Checklist and PBR users found different *types* of defects. The data led to inconclusive results on adequacy of the techniques for specific types of defects. For example, for defects of type ‘ambiguous’ (information

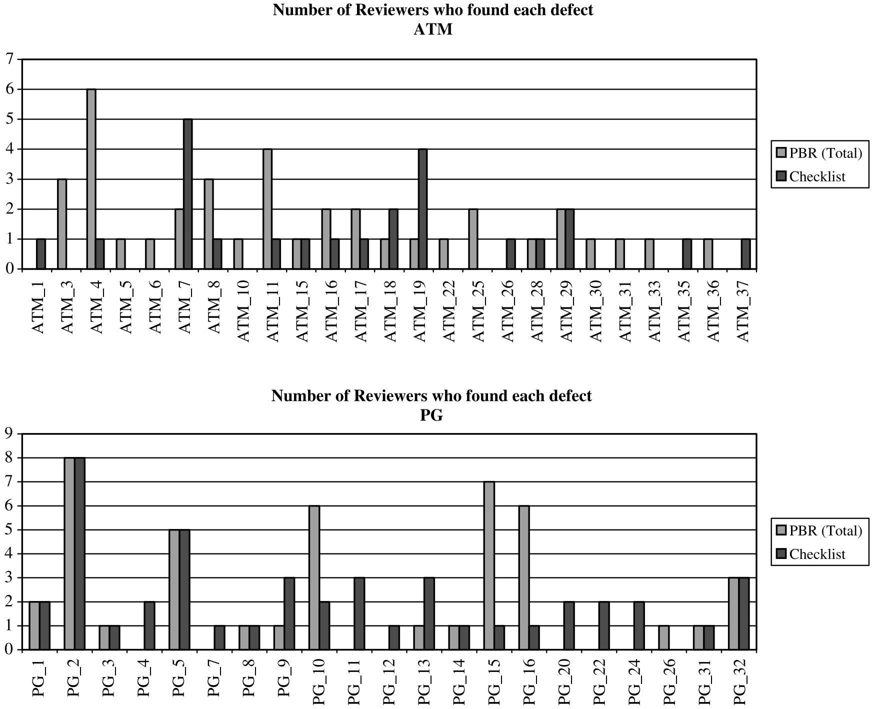

**Fig. 5** Distribution of number of reviewers detecting each defect

is not specified clearly enough and could be interpreted in multiple ways), the subjects using PBR were more effective than the checklist subjects in the ATM, but less effective in the PG. The information in Fig. 5 shows that some defects were found more often by users of one technique or the other (e.g., ATM_4 was found by 6 PBR users and only 1 checklist user). After examining this set of defects we were not able to find any obvious relationships among those defects that were found more often by PBR or those that were found more often by checklist. Due to the small number of defects that fell into these categories, we did not want to speculate on any spurious relationships at this point. Further study will be required to understand this effect. The suitability of a particular technique for a defect type is an interesting point to be addressed in further studies:

1) Is there a defect type for which one of the techniques would be more effective?
2) Does each technique produce uniform results such that a majority of reviewers using that technique identify defects of a particular type? What about for specific PBR perspectives?

### *3.5.5. R1: Do the PBR Perspectives Have the Same Effectiveness and Efficiency?*

Figure 6 shows plots of individual effectiveness and efficiency of the inspectors grouped by the document and perspective. Each point represents the mean of the 3 reviewers composing the group. The reviewers using the Designer perspective were the most effective and efficient on the ATM document. On the PG document, the reviewers using the Tester perspective were the most effective while the reviewers using the Designer perspective were the most efficient. The data from both documents was combined and an ANOVA was run to test whether the perspectives had a significant effect on either effectiveness or efficiency. The results of that test, shown in Fig. 7, indicate no significant influence on the effectiveness or efficiency (p = .654, p = .182). Based on this data, we can say that the effectiveness and efficiency of the perspectives were similar.

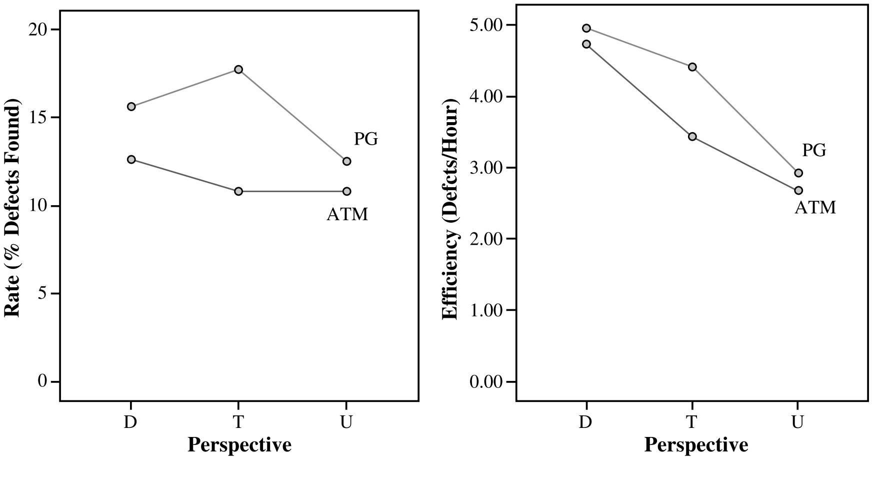

**Fig. 6** Rate and efficiency by perspective

**ANOVA Table for Rate (% Defects Found)**

| | DF | Sum of Squares | Mean Square | F-Value | p-value |
|---|---|---|---|---|---|
| Document | 1 | .007 | .007 | 2.29 | .156 |
| Perspective | 2 | .003 | .001 | .438 | .655 |
| Document x Perspective | 2 | .002 | .001 | .374 | .696 |

**ANOVA Table for Efficiency (Defects/hour)**

| | DF | Sum of Squares | Mean Square | F-Value | p-value |
|---|---|---|---|---|---|
| Document | 1 | 1.032 | 1.032 | .325 | .579 |
| Perspective | 2 | 12.528 | 6.264 | 1.973 | .182 |
| Document x Perspective | 2 | .557 | .278 | .088 | .917 |

**Fig. 7** ANOVA for rate and efficiency (by perspective)

### *3.5.6. R2: Do the PBR Perspectives Find Different Defects?*

This question asks whether the perspectives complement each other, in terms of defects found. If the perspectives are complementary, then a benefit is gained from using the entire collection, although it may be more expensive to use multiple reviewers.

As in the original study, we present Venn-diagrams in Figs. 8 and 9 indicating the amount of overlap among users of the three perspectives. For these figures:

- Part (a) shows how many defects were found by at least one user of each perspective and the total number of defect occurrences found by all users of each perspective (in parenthesis), e.g., in Fig. 8, six defects were found only by Designers and one defect was found by both Designers and Testers. The users of the Designer perspective found 11 different defects (6 + 1 + 2 + 2) and 14 total occurrences.
- Part (b) categorizes the defects based on which perspective found the greatest number of occurrences, e.g., in Fig. 8, seven defects were reported more times by Designers than by Testers or Users.

For the ATM document, the perspectives appear to be complementary. Figure 8a shows that each perspective identified unique defects with little overlap. The perspectives identified a similar number of occurrences overall (Designer—14,

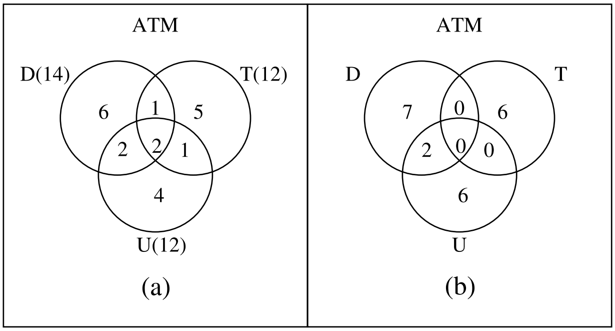

**Fig. 8** Number of defects found by each perspective, for ATM document

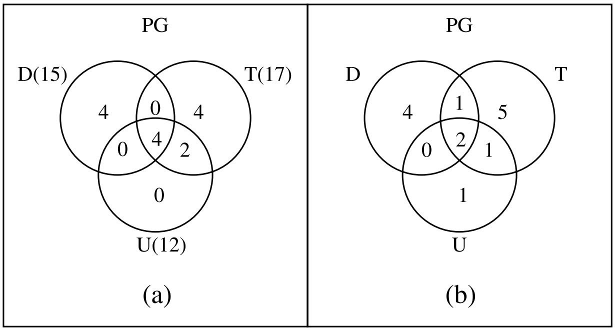

**Fig. 9** Number of defects found by each perspective, for PG document

Tester—12, and User—12). In terms of the perspective that was most effective at finding each defect, Fig. 8b shows that users of the three perspectives were more likely to find different defects. For the PG document, the Designer and Tester perspectives appear to be complementary, but the User perspective does not provide much added benefit. The perspectives again identified a similar number of occurrences (Tester —17, Designer—15, and User —12). In terms of defects found, Fig. 9a shows that inspectors applying the Designer and Tester perspectives found defects that were not found by the other perspectives, but those applying the User

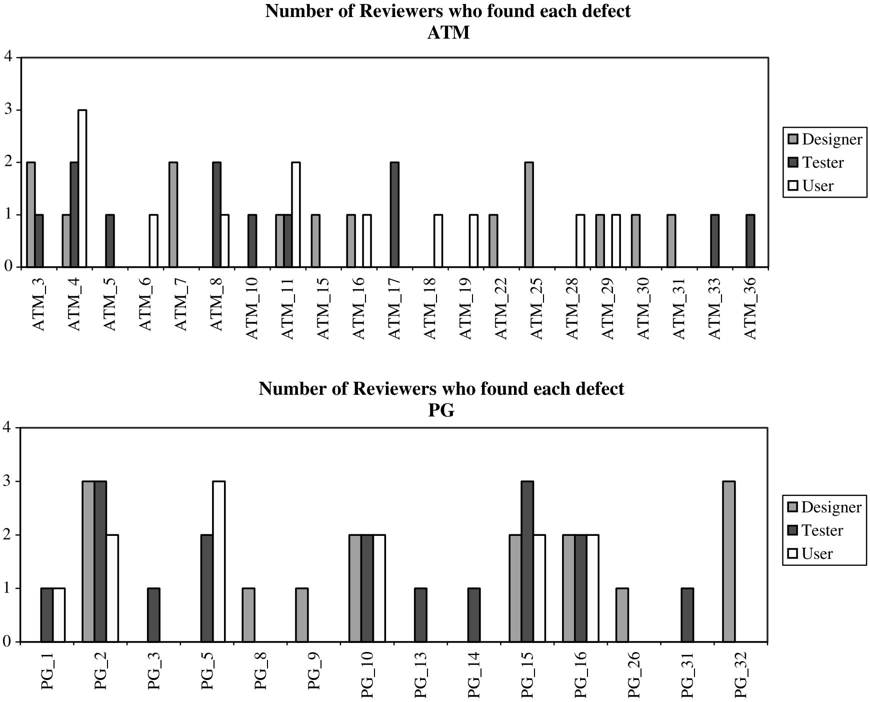

**Fig. 10** Distribution of number of reviewers that found each defect

| **ATM Document** | Correlation | p-value | **PG Document** | Correlation | p-value |
|---|---|---|---|---|---|
| User, Tester | .260 | .256 | User, Tester | .827 | .01 |
| User, Designer | -.084 | .716 | User, Designer | .315 | .272 |
| Tester, Designer | -.275 | .227 | Tester, Designer | .307 | .286 |
| | **21 Observations** | | | **14 Observations** | |

**Fig. 11** Pearson Correlation analysis of the perspectives for each document in the replication

perspective did not find any unique defects. None of the other variables we collected indicated why this result occurred (e.g., User reviewers were not less experienced than the other groups.)

The bar charts in Fig. 10 show the number of reviewers from each perspective that identified each defect (defects that were found by no inspectors are excluded). We can see from this data that for some defects one perspective outperformed the other perspectives (e.g., ATM_17, ATM_25, and PG_32). This result suggested that the perspectives may be complementary. To investigate this result further, we ran a Pearson's Chi Square analysis as done by Regnell et al. to test for differences in the distribution of the defects found by the different perspectives. While conducting the Chi Square test it became apparent that the data did not meet the requirements for that test. Specifically, the expected values (which are computed as the number of observations expected for each category) were all less than 5 and most were less than 1, violating one of the underlying assumptions of the test (Robson, 2002). Therefore, we have chosen to exclude the Chi Square analysis from this discussion.

Finally, in order to understand the degree of correlation between the perspectives (that is, if a subject from one perspective found a defect, was it likely that subjects from other perspectives also found the defect?), we conducted a correlation analysis similar to the analysis done by Regnell et al. The correlation analysis compares the sets of defects found by each perspective to determine how closely related they are. Figure 11 presents the results of that analysis with the Pearson correlation coefficient. The results in Fig. 11 show that out of the six possible pairwise comparisons, only one showed a significant correlation. In all other cases, the perspectives were not significantly correlated to each other. This result allows us to conclude, in contrast to the conclusion drawn by Regnell et al., that in general the subjects using different perspectives find different defects.

To further investigate the discrepancy between our results and the results presented by Regnell et al., we chose to reanalyze the data from the original study using these more sophisticated statistical tests. Again the assumptions for the Chi Square test were not met, so that analysis is excluded. The results of the correlation analysis done using the Pearson coefficient appear in Fig. 12. The results from the

| **ATM Document** | Correlation | p-value | **PG Document** | Correlation | p-value |
|---|---|---|---|---|---|
| User, Tester | -.083 | .578 | User, Tester | .075 | .63 |
| User, Designer | .155 | .298 | User, Designer | -.008 | .96 |
| Tester, Designer | .105 | .483 | Tester, Designer | .099 | .521 |
| | **47 Observations** | | | **44 Observations** | |

**Fig. 12** Pearson Correlation analysis of the perspectives for each document in the original study

original study support the results obtained in our replication. That is, there are no significant correlations among the perspectives.

### 3.6. Threats to Validity

Due to the nature of software engineering experimentation, the threats to validity of the results of studies must be identified and addressed. Because this was the first experiment conducted by the replicators, it is possible that experimental protocols were not followed as closely as necessary for accurate results. In order to combat this threat, the replicators took two specific actions:

1) Interaction with the original experimenters—Through this interaction, the replicators were able to have questions answered and doubts clarified.
2) Running a pilot study—The pilot study allowed the replicators to better understand how to run an experiment and debug their techniques and procedures.

In addition, there were some internal and external threats to validity of this experiment, which are discussed below.

*Internal Threats to Validity*

The first threat to internal validity was concerned with *testing*, whether bias could be introduced due to the way data was collected. The original study was run in the USA where the subjects either spoke English as their native language, or used English on a daily basis. In the replication, the subjects were from Brazil, so they did not speak English as their native language nor did they work in English on a daily basis. The class lecture notes, assignment instructions, techniques and artifacts were all written in English, so the lack of proficiency in English could have affected the results of the study.

A second threat to internal validity is *learning*. Based on the design in Fig. 2, it can be seen that the same subjects were trained and performed both the ‘checklist inspection’ and the ‘PBR inspection.’ Therefore, it is possible that the performing the ‘checklist inspection’ had some effect on performing the ‘PBR inspection.’ The study was designed in this way to combat a more serious threat to validity. PBR is a more procedurally defined technique than a checklist. If subjects were trained in the use of the more procedural PBR techniques prior to using the less-procedural checklist, there was a danger that they would perform the tasks on the checklist in a more ordered fashion that would be normally seen.

A second threat to internal validity concerned conformance to the original experimental protocols. There were some changes made to the experimental procedures by the replicators before running the study. There are two main issues that must be considered for this threat:

1) The replicators made some adjustments to the training time but kept it equal for both techniques. Although equal time was used in the original study, it probably is not fair to use equal training time for two techniques of varying complexity, since they provide different levels of detail and require different levels of background knowledge. As mentioned before, for each PBR perspective, the operational aspects have to be taught in addition to the understanding of the questions and of the required defect taxonomy.

2) The techniques were applied just after training, without giving the subjects time to mature and assimilate the underlying concepts. Therefore, it is possible that the subjects did not fully understand or completely follow the techniques during the experiment.

A third threat to internal validity is the process conformance of the subjects. We did not observe the subjects while they were working nor did we collect any intermediate artifacts. Therefore, we cannot be certain that the subjects followed the techniques that they were assigned. We have not reason to believe that they did not follow the techniques but we cannot verify this assumption.

*External Validity*

Because this study was run in the classroom at a university, the subjects were not as experienced as industrial professionals. Based on the data collected about the subjects' backgrounds, it was clear that most subjects were inexperienced in their PBR perspective. As a result, the conclusions drawn from this study may not be directly transferable to industrial inspectors. A second threat to external validity is the small number of subjects who participated in the replication. It is possible that any result seen here is a function of this small sample size.

## 4. Interpretation of Results

This paper described two runs of a study comparing PBR to other techniques for guiding software inspections. The overall results were not conclusive: In the original study, PBR was significantly more effective (at the 0.05-level) for both requirements documents on which it was applied. In the replication, there were no statistically significant results for either document, although a non-significant improvement was seen in one case. Looking more specifically at the individual PBR perspectives, we saw an interesting dichotomy: In the original study there was no significant correlation between any perspectives (i.e., they each found different sets of defects) while in the replication, two perspectives were significantly correlated on the PG document, which was the only document where no improvement was observed for PBR, significant or otherwise. Therefore, we can now hypothesize that for PBR to be more effective, the individual perspectives must effectively target different classes of defects. This implies that more attention must be given to issues such as:

- How representative the perspectives are of the real stakeholders of the document.
- How much experience the users of the perspectives have.

Unfortunately, the Regnell study which found little difference among the individual perspectives did not include comparison to a baseline inspection approach. Therefore, we cannot analyze whether or not it supports our new hypothesis. However, these results do point to a need for understanding under which conditions the individual perspectives can in fact be effectively focused on different defect types, thus contributing to the overall effectiveness of the entire family of techniques.

## 5. Conclusions and Future Plans

This paper described a replication of a study conducted on PBR. The results showed that PBR was more effective than Checklist for one of the two requirements documents used. The results also showed that for the same document on which PBR was more effective, the PBR and Checklist techniques were complementary. A reviewer's experience in the PBR perspective appeared to have little impact on his or her effectiveness. In terms of perspectives, there was not a large variation in the effectiveness (i.e., the overall number of defects detected) of the three perspectives overall.

Perhaps more importantly, however, these studies provided more valuable insight into whether the defects found by the various perspectives were the same or different and the impact that difference (or lack thereof) had on the overall effectiveness. While in most cases the perspectives turned out to find different defects (i.e., they were not correlated), in the case where the perspectives found similar defects (i.e., correlated) there was no overall benefit observed for the perspective-based technique. These results lead to important hypotheses about the particular aspects of the perspective-based approach that may lead to the overall benefit that has been observed.

A concrete result of this replication is that the Laboratory Package, which serves as a baseline for further experimental replications, has been evolved. Changes to the laboratory package included updating the list of defects to include new defects found in the replication, creating a list of frequent false positives, and evolving the feedback questionnaire to capture information on the cultural aspects of the subjects. In this questionnaire the subjects were asked to discuss whether the current list of defects was reasonable. This important information impacts the global analysis of the experiments, and will facilitate future efforts to analyze data across different replications. This laboratory package is available at the Readers' Project homepage http://www.labes.icmc.usp.br/readers).

A point for further investigation is the complementary aspects of PBR and Checklist, since both lead to finding some unique defects and finding some common defects. This investigation should cover both the types of defects uncovered by each technique in each document, and the number of defect occurrences. Such information is relevant to analyze the uniformity and adequacy of these techniques for uncovering specific types of defects, and to provide hints for improving them by establishing reading and inspection strategies tailored to different situations. This point will be addressed in forthcoming papers based on further replications.

Analyzing the results of this type of experiment is not an easy task, so we are exploring the use of Visual Data Mining (VDM) (Mendonca and Sunderhaft, 1999). VDM brings the possibility of better exploration and understanding of the results. With VDM, we can execute a discovery-driven, as opposed to hypothesis-driven, data analysis. This approach allows the many intervening factors that can significantly affect the results to be more fully explored.

**Acknowledgments** We would like to thank the members of the Readers' Project team for their comments during the project workshops, which helped improve this paper. We would also like to thank the subjects who took part in the studies described here. Finally, we would like to thank the anonymous reviewers whose comments helped us to improve this paper. This work has been supported by CNPq and by NSF (CCR9900307 and CCR0086078, establishing the Center for Empirically Based Software Engineering).

## References

Basili V, Green S, Laitenberger O, Shull F, Sorumgaard S, Zelkowitz M (1996) The empirical investigation of perspective based reading. Empir Softw Eng-Int J. 1:133–164

Ciolkowski M, Differding C, Laitenberger O, Munch J (1997) Empirical investigation of perspective-based reading: a replicated experiment. ISERN 97–13

Conradi R, Mahagheghi P, Arif T, Hegde LC, Bunde GA, Pedersen A (2003) Object-oriented reading techniques for inspection of uml models—an industrial experiment. Proc of Eur Conf on Object Oriented Program (ECOOP'03). Darmstadt, Germany, pp. 483–500

Laitenberger O, Atkinson C, Schlich M, El Emam K (2000) An experimental comparison of reading techniques for defect detection in uml design documents. J Syst Softw 53: 183–204

Laitenberger O, El Emam K, Harbich TG (2001) An internally replicated quasi-experimental comparison of checklist and perspective based reading of code documents. IEEE Trans. Softw Eng 27:387–421

Mendonca M, Sunderhaft NL (1999) A state of the Art Report: Mining Software Engineering Data. Department of Defense (DoD) Data & Analysis Center for Software Engineering Data

Porter A, Votta L (1998) Comparing detection methods for software requirements inspections: a replication using professional subjects. Empir Softw Eng-Int J 3:355–379

Regnell B, Runeson P, Thelin T (2000) Are the perspectives really different? Further experimentation on scenario-based reading of requirements. Empir Softw Eng 5:331–356

Robson C (2002) Real World Research. Blackwell Publishing, Malden: MA

Shull F (1998) Developing Techniques for Using Software Documents: A Series of Empirical Studies. Ph.D. Thesis. Department of Computer Science, University of Maryland, College Park

Shull F, Basili V, Boehm B, Brown, AW, Costa P, Lindvall M, Port D, Rus I, Tesoriero R, Zelkowitz M (2002a). What we have learned about fighting defects. Proc IEEE Symp on Softw Metr 249–258

Shull F, Basili V, Carver J, Maldonado J, Travassos G, Mendonca M, Fabbri S (2002b) Replicating Software Engineering Experiments: Addressing the Tacit Knowledge Problem. Proc of Int Symp Empir Softw Eng (ISESE'02). Nara, Japan, pp. 7–16

Zhang Z, Basili V, Shneiderman B (1999) Perspective-based usability inspection: an empirical validation of efficacy. Empirical Software Engineering—An International Journal 4:43–70

**José Carlos Maldonado** received his BS in Electrical Engineering/Electronics in 1978 from the University of São Paulo (USP), Brazil, his MS in Telecommunications/Digital Systems in 1983 from the National Space Research Institute (INPE), Brazil, and his D.S. degree in Electrical Engineering/Automation and Control in 1991 from the University of Campinas (UNICAMP), Brazil. He worked at INPE (the Brazilian Space Agency) from 1979 up to 1985 as a researcher. In 1985 he joined the Departamento de Ciências de Computação of the Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP) where he is currently a faculty member. His current research interests include: software quality; software testing, debugging and maintenance; object orientation and aspects; software metrics; and training and education. He is a member of the SBC (the Brazilian Computer Society), ACM and of the IEEE Computer Society.

**Jeffrey Carver** is an Assistant Professor in the Computer Science and Engineering Department at Mississippi State University. He received his PhD from the University of Maryland in 2003. His PhD thesis was entitled "The Impact of Background and Experience on Software Inspections." His research interests include: Empirical Software Engineering, Software Inspections, Qualitative Methods, Software Process Improvement, Software Metrics, and Software Engineering for High Performance Computing.

**Forrest Shull** is a scientist at the Fraunhofer Center for Experimental Software Engineering, Maryland (FC-MD). He received his doctorate degree in Computer Science from the University of Maryland, College Park in 1998. At FC-MD he is project manager and member of the technical staff for projects with clients that have included Fujitsu, Motorola, NASA, and the U.S. Department of Defense. He is responsible for basic and applied research projects in the areas of software inspections, software defect reduction, general technology evaluation, and software measurement. A primary focus of his work has been developing, tailoring, and empirically validating improved techniques for inspections of software artifacts, including requirements and design documents.

**Sandra Fabbri** is an assistant professor in the Department of Computer Science at Federal University of São Carlos, São Paulo, Brazil. She has a degree in Computer Science from Unicamp–University of Campinas. She received a MSc in Information System area and a PhD in Software Engineering area from USP–University of São Paulo. Her research interests include Software Testing, Formal Specification, Software Quality, Requirement Engineering and Experimental Software Engineering.

**Emerson Silas Dória** received the degree in computer science from Universidade do Oeste Paulista, Presidente Prudente, SP, in 1995, and the M.S. degree in computer sciences and computational mathematics from Universidade de São Paulo, São Carlos, SP, in 2001. He is currently Professor at the Faculdade de Informática de Presidente Prudente, Universidade do Oeste Paulista. He is involved with software development process, analysis and design techniques for object-oriented system development. Msc. Emerson is member of the Brazilian Computer Society.

**Luciana Martimiano** received her BS in Computer Science in 1995 from the State University of Maringá (UEM), Brazil, her MSc in Computer Science in 1999 from the Institute of Mathematical Sciences and Computing at the University of São Paulo (ICMC-USP), Brazil, and her PhD is in progress in the Institute of Mathematical Sciences and Computing at the University of São Paulo (ICMC-USP), Brazil. She is currently at the Computer Science Department of the Institute of Mathematical Sciences and Computing at the University of São Paulo (ICMC-USP), where she is a faculty member since 2002. Her research interests are in the areas of Information Security Management, Ontologies applied to Security Domain, and Security Data Standardization.

**Manoel G. Mendonça** received his Ph.D. in computer science from the University of Maryland in 1997. He also holds a M.Sc. in computer engineering from the State University of Campinas (1990), and a B.S.E.E. in electrical engineering from the Federal University of Bahia (1986), both in Brazil. Dr. Mendonca was a visiting scientist and was awarded a doctoral fellowship from the IBM Toronto Laboratory's Centre for Advanced Studies. Between 1997 and 2000, he worked as a Faculty Research Associate at the University of Maryland and as a scientist at the Fraunhofer Center for Experimental Software Engineering in Maryland. He is now a professor at Salvador University (UNIFACS) in Brazil, where he heads the Computing Research Center (NUPERC). His main research interests are on information visualization, experimental software engineering, and knowledge management information systems.

**Dr. Victor R. Basili** is a Professor of Computer Science at the University of Maryland. He was founding director of the Fraunhofer Center for Experimental Software Engineering, Maryland, and one of the founders of the Software Engineering Laboratory (SEL) at NASA/GSFC. He received a B.S. from Fordham College, an M.S. from Syracuse University, and a PH.D. in Computer Science from the University of Texas at Austin. He has been working on measuring, evaluating, and improving the software development process and product for over 30 years. Methods for improving software quality include the Goal Question Metric Approach, the Quality Improvement Paradigm, and the Experience Factory organization.