Instituto de Ciências Matemáticas e de Computação

ISSN:

DATA ANALYSIS OF THE FIRST TWO PBR REPLICATIONS IN READERS PROJECT

Sandra Fabbri

DC-UFSCar
sfabbri@dc.ufscar.br

José C. Maldonado
Erika Nina Höhn
Maria C. F. de Oliveira
Luciana Martimiano

ICMC-USP
{jcmaldon, hohn, cristina, luciana}
{@icmc.usp.br}

Manoel Mendonça

UNIFACS
mgmn@unifacs.br

Forrest Shull[1]
Jeff Carver[3]
Victor Basili[1,2]

[2]Fraunhofer Center
[1]University of Maryland
[3]Mississippi State University
fshull@fc-md.umd.edu,
carver@CSE.MsState.EDU,
basili@cs.umd.edu

No.: TR-225

RELATÓRIOS TÉCNICOS DO ICMC

São Carlos, 2004

Contents

List of Figures iii

List of Tables iv

Abstract v

1. Introduction 1

2. The PBR Technique: Related Work 2

3. The Replications 4

3.1 The First PBR Replication (R1) 5

3.2 The Second PBR Replication (R2) 6

3.3 Subject Profile and Motivation: R1 and R2 7

4. Comparative Data Analysis 9

4.1. OS1' Do PBR teams detect more defects than Checklist teams? 9

4.2. OS2' Do individual PBR or Checklist reviewers find more defects? 9

4.3. OS3' Does the reviewer's experience affect his or her effectiveness? 15

4.4. RS1 Do individual reviewers using PBR and Checklist find different types of defects? 18

4.5. RS2 Do the PBR perspectives have the same effectiveness? 25

4.6. RS3 Do the PBR perspectives find different types of defects? 26

4.7. Feedback from Subjects 30

5. Threats to Validity 30

6. Summary, Insights and Future Work 31

References 33

Annex A 35

List of Figures

Figure 1 – Families of Reading Techniques 3

Figure 2 – Summary of Subject Profiles of Replications 1 and 2 8

Figure 3 – Performance for both replications 12

Figure 4 – Subjects individual performance regarding the percentage of defects found: ATM (Checklist) and PG (PBR). 13

Figure 5 – Subjects individual performance regarding the percentage of defects found: PG (Checklist) and ATM (PBR) 13

Figure 6 – R1: Individual subject efficiency 14

Figure 7 – R2: Individual subject efficiency 14

Figure 8 – PBR effectiveness versus readers' role experience (R1) 15

Figure 9 – PBR effectiveness versus readers' role experience (R2) 16

Figure 10 – PBR effectiveness versus readers' role experience (R1+R2) 16

Figure 11 – Checklist effectiveness versus reader's role experience (R1) 17

Figure 12 – Checklist effectiveness versus reader's role experience (R2). 17

Figure 13 – Checklist effectiveness versus reader's role experience (R1+R2) 18

Figure 14 – (a) ATM: defects found per technique; (b) PG: defects found per technique 19

Figure 15 – (a) ATM: defects found per technique; (b) PG: defects found per technique 19

Figure 16 – ATM-Checklist - Types of defects found by each subject. 23

Figure 17 – ATM-PBR –Types of defects found by each subject. 23

Figure 18 – PG-Checklist - Types of defects found by each subject. 24

Figure 19 – PG-PBR - Types of defects found by each subject. 24

Figure 20 – R1/ATM (a) Identification of the different defects found by the perspectives; (b) Grouping the defects by perspective that obtained better or equal performance than the other perspectives. 26

Figure 21 – R1/PG (a) Identification of the different defects found by the perspectives; (b) Grouping the defects by perspective that obtained better or equal performance than the other perspectives. 27

Figure 22 – R2/ATM (a) Identification of the different defects found by the perspectives; (b) Grouping the defects by perspective that obtained better or equal performance than the other perspectives. 28

Figure 23 – R2/PG (a) Identification of the different defects found by the perspectives; (b) Grouping the defects by perspective that obtained better or equal performance than the other perspectives. 28

Figure 24 – R1+R2/ATM (a) Identification of the different defects found by the perspectives; (b) Grouping the defects by perspective that obtained better or equal performance than the other perspectives. 29

Figure 25 – R1+R2/PG (a) Identification of the different defects found by the perspectives; (b) Grouping the defects by perspective that obtained better or equal performance than the other perspectives. 29

List of Tables

Table 1– Taxonomy for Defects on Requirements Documents 3
Table 2 – ANOVA summary table with relation to the individual effectiveness 10
Table 3 – ANOVA summary table regarding individual subject efficiency 10
Table 4 – Comparing results for both requirements documents and both replications 11
Table 5 – Pearson and Spearman's correlation coeficients – PBR effectiveness 15
Table 6 – Pearson and Spearman's correlation coeficients – Checklist effectiveness 16
Table 7 – ATM - Checklist/PBR: Percentage of defects found by defect type 20
Table 8 – ATM - Checklist/PBR: Percentage of defect occurrences by defect type 20
Table 9 – ATM - PBR: Number of defects found per perspective 20
Table 10 – ATM - PBR: Number of defects occurences per perspective 21
Table 11 – PG - Checklist/PBR: Percentage of defects found by defect type. 21
Table 12 – PG - Checklist/PBR: Percentage of defect occurrences by defect type 21
Table 13 – PG - PBR: Number of defects found per perspective 22
Table 14 – ATM - PBR: Number of defects occurences per perspective 22
Table 15 – ATM/PG - PBR: R1 Average percentage of defects found and defect observation 25
Table 16 – ATM/PG - PBR: R2 Average percentage of defects found and defect observation 25
Table 17 – ATM/PG - PBR: R1+R2 Average percentage of defects found and defect observation 26
Table 18 - Subjects comments about the experiment. 30

Abstract

This Technical Report describes two empirical studies carried out in the context of the Project "Readers: A Collaborative Research to Develop, Validate and Package Reading Techniques for Software Defect Detection" – where Brazilian and American researchers investigate the effectiveness of software requirements documents inspection techniques under diverse technical and cultural settings. The studies conducted are replications of a previous experiment on a family of reading techniques named PBR – Perspective Based Reading, and had as subjects undergraduate students enrolled in computing courses at ICMC/USP and DC/UFSCar. Identified as replications R1 and R2, respectively, they compared the PBR and Checklist techniques for Requirements Documents analysis. Four metrics were used to evaluate the data collected: Defects Found, Occurrences of Defects, Effectiveness and Efficiency. Although both replications produced similar results for one of the two requirements documents used, some conflicting results were produced for the other document. In R1, PBR did better than the Checklist technique on one of the documents, in agreement with results from previous studies. However, for the other document, Checklist did better in terms of the subject's effectiveness, one of the metrics applied for analyzing the results. These conflicting results are discussed, possible sources of variation amongst the experiments are identified and actions to mitigate such problems in future replications are proposed. Neither Checklist nor PBR led to complete uniformity of defect reporting, but with PBR a higher percentage of subjects achieved the same higher performances (within each perspective) in both replications.

Keywords:
software engineering experimental replication, laboratory package, reading techniques, requirements documents, PBR

1. Introduction

Software Engineering still has to evolve from a discipline that simply provides assertions about the effects of a technique into a scientific discipline based upon observation, theory formulation and experimentation. Seeking this goal, many researchers conduct empirical studies to evidence the quality and productivity of software development methods, techniques and tools [Basili1996; Fusaro1997; Lott1997; Porter1995, Regnell2000].

Empirical research is crucial, but experience has shown that it is extremely difficult to build a usable body of knowledge from isolated studies. Accepting results from a single experiment on a topic as the final word without considering differences in system domains, subject profiles and cultural environments may be a gross mistake. Empirical research should not be concerned just with running individual studies but rather with enhancing the understanding of software development processes, the costs and benefits of classes of techniques and, ultimately, consolidating a body of knowledge and establishing novel software development models. It is therefore imperative to run more studies and to search for an integrated framework to support the analysis of the whole body of results obtained.

Producing and integrating a significant body of results from controlled experiments on families of technologies can only be achieved through collaborative work. The problem of conducting effective replications is addressed in a cooperative project initiated in 1999 involving Brazilian and American researchers, named “Readers: A Collaborative Research to Develop, Validate and Package Reading Techniques for Software Defect Detection”. Supported by the Brazilian (CNPq) and American (NSF) national research funding agencies, this project investigates techniques for software document analysis in diverse technical and cultural settings [Maldonado2002].

Within its scope replications were conducted of previous experiments designed to study the application of human-based review techniques to find defects in software requirements documents [Basili1996; Fusaro1997]. The focus on reviews and underlying reading techniques is justified by their relevance since most software development documents require continual understanding, review, and modification throughout the development life cycle. A long string of studies has demonstrated the effectiveness of techniques for improving individual review practices in different domains and types of inspection: requirements tailored to natural language [Basili1996], formal notation [Porter1995], high-level designs [Laitenberger2000a; Shull2001], code [Basili1987, Laitenberger2000b], and user interfaces [Zhang1999]. In the context of the Readers Project, the empirical studies replicated compare reading techniques for Requirements Document analysis, in particular PBR – Perspective Based Reading – with Ad-Hoc or Checklist approaches. Though previous comparisons have already been conducted, many questions remain open to further investigation:

- Do PBR teams detect more defects than Checklist teams?
- What is the impact of a reviewer’s experience on his effectiveness when using PBR?
- Do the PBR perspectives differ in terms of effectiveness and specific types of defects found?
- Should the level of detail in a technique vary according to the experience of the reviewer?
- Can PBR be tailored for different development approaches (e.g. waterfall vs. spiral lifecycle models)?
- Is PBR useful on various types of software (e.g. middleware vs. user interface software)?

In this Technical Report we tackle the first three issues above analyzing results and discussing insights from two initial replications of an experiment [Basili1996; Fusaro1997] designed to verify the improvement in effectiveness brought from the use of PBR over a typical Checklist approach. More than just verifying hypotheses raised by the original experiment, these replications provided a framework for more comprehensive studies on [Shull2002, Maldonado2002]:

- Generating and facilitating experimental collaborations.
- Transferring know-how on the execution of experiments and replications.
- Exploring new data analysis methodologies.
- Packaging experimental artifacts.

Although both replications produced similar results for one of the two requirements documents inspected – PBR did better than the Checklist technique, in agreement with previous results [Basili1996; Fusaro1997; Shull2001] –, some conflicting results were produced for the second document in the first replication, where Checklist did better on one of the analysis metrics. We discuss these conflicting results, identifying possible sources of variation amongst the experiments and proposing actions to mitigate such problems in future replications. Our ultimate goal is to contribute to the high-level issues mentioned previously.

The remainder of this text is organized as follows. In Section 2 we discuss the PBR technique and related work investigating its quality and productivity. In Section 3, we describe how the two Readers replications of the original PBR experiment were conducted. In Section 4 we present and compare the results from these replications. Finally, in Section 5, insights, conclusions and directions for further research are presented.

2. The PBR Technique: Related Work

Perspective-Based Reading (PBR) is a family of reading techniques that guide a reader in looking for defects in a natural language Requirements Document. PBR was developed by the Experimental Software Engineering Group at the University of Maryland (Basili, 1996) – one of the partners in the Readers Project. PBR defines a series of perspectives, representing the major stakeholders of the requirements document. It provides the inspector with a process to assume one of those stakeholders perspectives and a set of instructions on how to read a software document (or artifact), or what to look for in order to uncover defects [Basili1996]. The “basic set” of perspectives was defined to be a software designer (D), a tester (T) and an end-user (U). Based on his perspective, an inspector creates an abstraction of the requirements relevant to that stakeholder. 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 is given a series of questions to help uncover defects. Questions are driven by a taxonomy of defects on requirements documents given in Table 1. This taxonomy is not assumed to be orthogonal or static, and it can be tailored to specific environments or domains.

Table 1- Taxonomy for Defects on Requirements Documents

Types
Ambiguous Information (A): Information within the requirements documents is inconsistent or ambiguous with other information in the document; **Inconsistent Information (II)**: Two sentences contained in the specification directly contradict each other; **Incorrect fact (IF)**: Some sentences contained in the requirements document/functional specifications assert a fact that cannot be true; **Extraneous Information (E):** Information is provided but is not needed or used; **Miscellaneous Defect (MD):** Other defects; **Omission (O)**: Necessary information about the system has been omitted from the requirements document.

Figure 1 shows PBR as one of several families of reading techniques developed for various purposes. Each family (and thus each technique) is associated with a particular document (e.g., requirements) and notation (e.g., Portuguese text, English text, or a formal notation). Each technique within a family is:

- Tailored, in that it is based upon a project and its environmental and cultural characteristics;
- Detailed, in that the reader must follow a well-defined set of steps;
- Specific, in that reading the document the reader has a particular goal and procedures that support this goal;
- Focused, in that it provides a particular coverage of the document, with a combination of techniques in the family providing a complete document coverage;
- Studied empirically to determine its effectiveness in different situations.

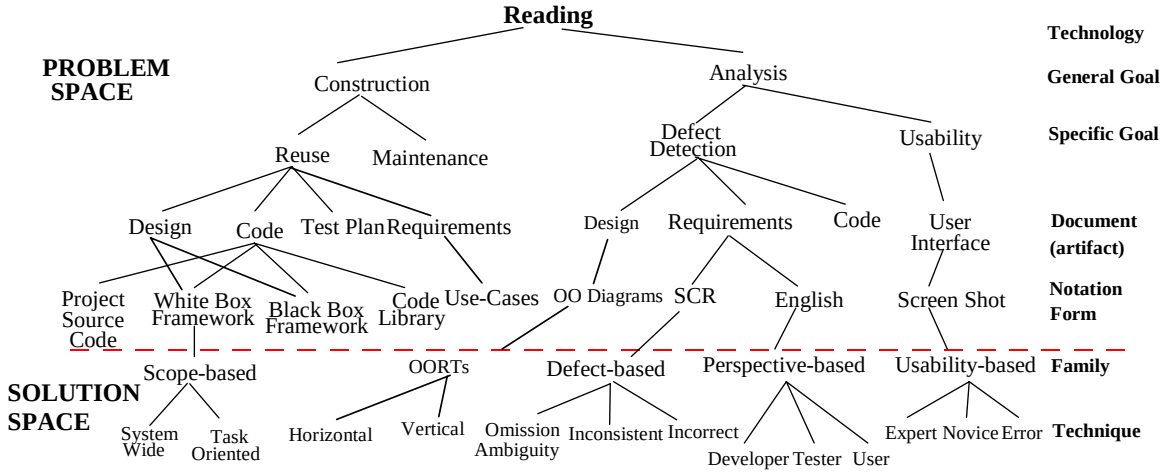

Figure 1 - Families of Reading Techniques

Since the first PBR experiments, it has been evaluated and improved empirically using over 150 software engineering students and 25 professionals from NASA Goddard Space Flight Center (Shull, 2000). In a summary of PBR experiments, Regnell et al. observe that results vary substantially (Regnell, 2000). Shull et al. argue that previous experiments provide evidence that PBR leads to improved effectiveness for both individual inspectors and inspection teams under certain conditions, e.g., when working with unfamiliar application domains. However, when working in familiar application domains, experienced inspectors

sometimes ignore the PBR procedure and use previously acquired heuristics. This fact suggests that PBR may be better suited for less experienced inspectors (Shull, 2000).

According to Shull et al. [Shull2002] existing studies on PBR show evidence of effectiveness, but further studies are necessary to refine such understanding into actionable heuristics. For example, studies on other populations are required to assess the impact of reviewers' experience on their effectiveness in applying the technique and to propose approaches for guaranteeing effective reviews by different classes of reviewers (the level of technique detail might possibly vary to match different experience levels). Such studies may also bring insight on how to tailor reading techniques to different existing practices, for example, waterfall versus spiral lifecycle models. Studies in different domains are also necessary to evaluate the suitability of techniques to different types of systems (e.g. middleware versus user interface software).

3. The Replications

The original experiment, run at the University of Maryland, compared the performance of teams of subjects using PBR and their usual reading technique for defect detection in software requirements [Porter1995]. This was a well-designed study whose treatments allowed multiple variables to be analysed and that provided some solid evidence that PBR was effective for inspection teams. The results of the original experiment were as follows:

- Teams of subjects using PBR found in overall more defects than teams using their usual technique. This result was statistically significant.
- Individual subjects inspecting two generic documents found more defects when using PBR than when using their usual technique. This result was also statistically significant. When inspecting specific NASA documents, individual subjects using PBR found slightly more defects than individual subjects using the usual technique. This result was not statistically significant.
- There was no consistent correlation between an inspector's experience in their PBR perspective and their inspection effectiveness.

A laboratory package has been organized aiming at building an experimental infrastructure for supporting future replications. A Laboratory package describes the 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. They establish a basis for confirming/denying original results, complementing the original experiment and tailoring the object of study to specific experimental contexts.

The design and experimental goals of the replications were refined to investigate additional variables, as suggested by the results commented above. In the replications the ´usual´ technique was replaced by a Checklist, and six questions were established for investigation: three were brought from the original study and three were derived from the open questions mentioned in Section 1.

- **Goals of the replication studies**

The three research questions from the Original Study, denoted OS1 – OS3, were:

OS1) If teams of individuals (such as during an inspection meeting) were giving unique PBR perspectives, would a larger collection of defects be detected than if each read the document in a similar way?
OS2) Would individuals reading a document using PBR find a different number of defects than if they used their 'usual' technique?
OS3) Does a reviewer's experience in his perspective affect his effectiveness with PBR?

The above questions were reformulated (OS1'–OS3') and three additional research questions have been addressed in the Replications Studies, denoted RS1–RS3:

OS1') Do PBR teams detect more defects than Checklist teams?
OS2') Do individual PBR or Checklist reviewers find more defects?
OS3') Does a reviewer's experience affect his or her effectiveness?
RS1) Do individual reviewers using PBR and Checklist find different defects?
RS2) Do the PBR perspectives have the same effectiveness?
RS3) Do the PBR perspectives find different types of defects?

A detailed description of the main steps of the replication process is given elsewhere [Maldonado2002]. Two requirements documents from the original study, containing 37 and 32 defects, respectively, were used in the replications: ATM, describing an Automated Teller Machine, and PG, describing the operation of a Parking Garage. Defects reported by subjects are either related to the list of defects from the original experiment, or are considered as new defects, or are taken as false positives. Both Checklist and PBR were applied in sessions of 01h45min each, though most subjects finished the reading activity before the allocated time had elapsed. In Replication 1 (R1) subjects were not asked to register the elapsed time between starting and finding/classifying each defect. In Replication 2 (R2) they were asked to register this information to allow further analysis on technique learning curve.

3.1 The First PBR Replication (R1)

In December 2000, 18 undergraduate students from the Software Engineering course at ICMC-USP carried out the experiment, which consisted of the following steps:

i) Subjects filled out the Consent and the Analyst Survey Forms and were assigned to one of two groups;
ii) Subjects applied the techniques as follows: On the first day, all subjects were trained in the baseline (Checklist) method. Subjects in Group 1 then reviewed the ATM Requirements Document and subjects in Group 2 reviewed the PG Requirements Document. On the second day, each subject was trained in one of the three PBR perspectives. Subjects then reviewed the other requirements document, i.e. Group 1 reviewed PG and Group 2 reviewed ATM. Experiment design is described in Annex A;
iii) Data was collected and results analyzed by experimenters; and
iv) Subjects received feedback from experimenters.

- Context:

Obtaining and using a Laboratory Package and its associated artifacts were key issues for this replication. Although the framework of the Readers Project ensured access to such a package, assembling a complete and consistent Lab Package for experiment replication is not an easy task. The original Lab Package included several artifacts that evolved over time as a result of conducting multiple replications. It was difficult to identify compatible and/or consistent artifacts due to version control and configuration issues arising from the growing number of artifacts in the UMD experience base – such issues are now being addressed by the CeBASE project [Basili2001a, Basili2001b, Goth2001], where knowledge management information systems are being used to assemble and manage a large body of knowledge on empirical software engineering. After gathering all the necessary artifacts, some of them still had to be adapted for the new replications. For example, questions had to be included in the Analyst Survey Form filled out by subjects to characterize their English expertise.

To handle such difficulties, several actions were taken by replicators. One of them was to seek close interaction with the original experimenters, in order to answer questions and clarify doubts. They also decided to run a Pilot Study to get a better understanding of the experimental process, including timing, tasks to be executed and documents to be delivered to subjects. The pilot study was important because none of the original experimenters would be present at the replication and tacit knowledge should be well understood. It helped the replication team to assess process conformance before undertaking any significant replication effort. The process for executing the replication experiment was carefully written down, documenting the timing and the entry and exit criteria for each step. Although not included in the original package, the team quickly discovered that this was key information for running replications with high process conformance in the absence the original experimenters [Dória2001].

Despite these precautions, the first replication produced conflicting results: PBR performed better than Checklist on the ATM document, a result in accordance with previous PBR experiments [Basili1996; Fusaro1997; Shull2001], but performed worse than Checklist on the PG document regarding effectivess, one of the analysis metrics. The reviewer's experience in their PBR perspective appeared to have little impact on their effectiveness. Also, there was no large variation in the effectiveness of the three perspectives overall, and they did appear to be complementary to each other in terms of defects uncovered.

3.2 The Second PBR Replication (R2)

In May 2001, 18 undergraduate students from the Software Engineering course at the Federal University of São Carlos conducted the second replication, which was quite similar to the first one, consisting of the following steps:

i) Subjects filled out the Consent and the Analyst Survey Forms;
ii) The experiment follows the experimental design of the previous ones, and was divided in four half-day periods. Subjects applied the techniques as follows: On the first half-day, all subjects were given an overview on inspection techniques and trained in the baseline (Checklist) method. In the second half-day subjects from Group 1 reviewed the ATM Requirements Document and subjects in Group 2 reviewed the PG Requirements Document. On the third half-day, each subject was trained in one of three PBR

perspectives. Then, in the fourth half-day, subjects reviewed the other requirements document, i.e., Group 1 reviewed PG and Group 2 reviewed ATM;

iii) Data Collection and Analysis of the Results by the Experimenters; and
iv) Subjects received feedback from the experimenters.

- Context:

Moving from the first replication (R1) to the second one (R2) was considerably simpler than moving from the UMD Lab Packages to the first replication, as the experiment material was essentially the same. Differences worth noting between both replications are in:

- Subject profile and motivation;
- Experiment distribution in time;
- Insertion of additional fields in the Defect Collection Form; and
- Trainer expertise.

Different persons trained subjects in both replications, and the trainer in R2 was more senior. Although in both cases it was their first training session, the trainer for R2 had closely followed the procedures in the first replication. Consequently, we believe that training sessions in R2 were as good as those of R1, if not better. The Defect Collection Form was slightly expanded to collect two additional attributes for each reported defect, namely the number of the requirement in which the defect appeared and the time of defect identification. Those attributes improved data analysis and reporting them requires almost no extra effort from subjects.

Observations similar to those made in R1 apply to R2, i. e., reviewer's experience in their PBR perspective appeared to have little impact on effectiveness. The complementary nature of the perspectives remained and some defects were found by only one perspective. In this replication, the number of defects found in common by the three perspectives, for the ATM document, was greater than in R1, and it was the same as R1 for the PG document.

It is worth noting that R1 was run in two consecutive full days, following the same procedure adopted in previous UMD PBR experiments [Basili1996; Shull2001]. During the feedback sessions, subjects observed that they could do better if allowed more time between the training and application sessions. We thus modified the procedure in replication R2: Checklist training was given in one day and Checklist was applied by subjects in the following day; a week later subjects were trained in PBR in one day and applied PBR in the following. Equal total times were spent in both replications. Controversially, in the feedback session of R2 subjects observed that they would rather have the experiment run in consecutive days, a clear indication that subjects are not necessarily reliable sources of feedback. This procedural change is a possible source of variation in the results, although we believe it is a minor one. On the other hand, we do believe that subject profile and motivation are major sources of variation in the results of this type of experiment. The following session discusses the role of such issues in our replications.

3.3 Subject Profile and Motivation: R1 and R2

Subjects in R1 were students taking a Software Engineering class. They were given a motivational presentation about the goals and activities of the Readers Project, but participation in the experiment was on a voluntary basis and deserved no extra credits. In R2 approximately two thirds of the subjects were taking the Software Engineering course for the

second time and were given credits for full participation in the experiment. The remaining third were student volunteers who had taken (and passed) an earlier edition of the same course and who also attended a motivational presentation. We may state that a third of the subjects in R2 had the same level of motivation of subjects in R1.

Subject profiles for both replications were organized, according to their assigned Group and PBR perspective, contemplating the following characteristics:

- English proficiency level in reading (**Q1**);
- Years of experience as Manager, Developer, Tester and Analyst (**Q2**);
- Years of experience in using requirements documents (**Q3**); and
- Years of experience in writing requirements documents (**Q4**).

In R1, subjects' English-language skills were sufficient, in overall (none inferior to a "medium" rating). Subjects in Group 1 were not significantly more experienced as software engineers than those in Group 2. The major difference on expertise is between subjects taking the User perspective in Group 1 (2 years as developers) and Group 2 (no experience as developers). Most subjects had no previous experience as managers, testers or analysts. About half had at least two years experience as developers and only two had more than three years experience. Thus, in general, subjects' industrial experience was low. Only one subject had significant previous experience (5 years as developer, 3 as tester, and 2 years as an analyst). Subjects were generally not highly experienced in their review perspectives; all PBR reviewers, except Designers, were more experienced as general developers than in their own perspective.

Likewise, in R2, subjects' English-language skills were sufficient (none inferior to a "medium" rating). Subjects in both Groups 1 and 2 had roughly the same level of experience as software engineers. The major difference on expertise is between subjects assuming the Designer perspective in Group 1 (on average, 2 years as developers) and Group 2 (on average, 0.5 year as developers). Most subjects had no previous experience as managers, testers, analysts, or developers. Only one subject in R2 had considerable experience (3.5 years as developer and 0.5 year as analyst). Except for one Designer and one Tester in Group 1, subjects were generally not highly experienced in their review perspectives.

Figure 2 summarizes the subjects' average years of experience for both replications. In general, subjects of R1 are more experienced as developers than those of R2, and experience as developers is also more evenly distributed in R1 than in R2. Moreover, in R1 both groups have equivalent experience, while in R2 Group 1 is slightly more experienced than Group2.

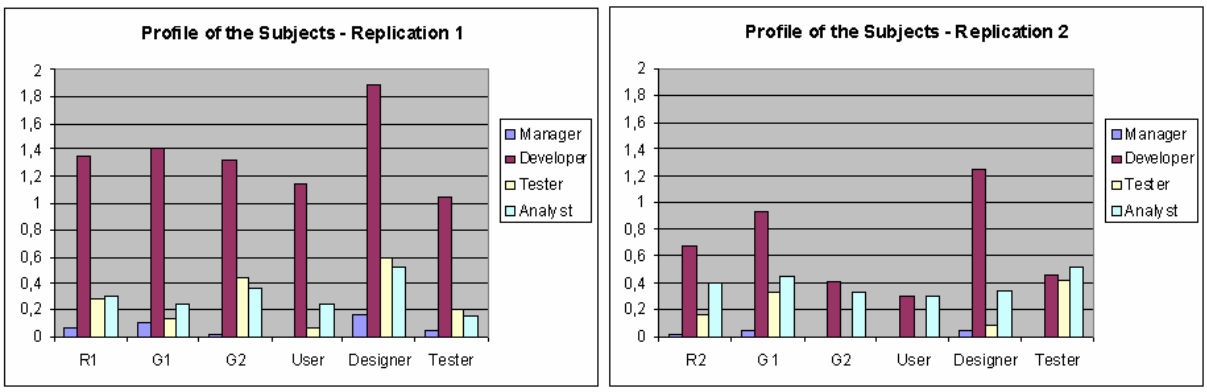

Figure 2 – Summary of Subject Profiles of Replications 1 and 2

4. Comparative Data Analysis

In this section we present the results collected in both replications, addressing the roll of questions listed in Section 3 as goals of the replication studies. Each question is commented regarding each individual replication, R1 and R2, and then considering the combination R1+R2. Since subject profiles were very similar in both replications, for the sake of data analysis R1+R2 has been considered as another replication whose data collected equals the data union from R1 and R2, giving us a greater data sample to work on.

Some hypotheses were formulated for analysis, based on the independent variables defined: the reading technique, the requirements document and the reader's experience. Isolating each variable and using a Testing Hypothesis to analyze the results we can verify the influence of the variables and determine whether the independent variables affects the results. Dependent variables were the individual subjects' effectiveness and efficiency. The statistical analysis was conducted using ANOVA, an Analysis of Variance Technique [Box1978], and MINITAB [Minitab2000] version 13.31. As the experimental design is balanced, ANOVA for balanced design was applied. The analysis is a 2X2 factorial experiment with repeated measures in blocks of size 2 [Winer1991], and involves two different factors, or treatments: the Reading Technique (RT) and the Requirements Document (DOC).

4.1. OS1' – Do PBR teams detect more defects than Checklist teams?

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

Ha: The defect detection rates of teams applying PBR are higher compared to teams using Checklist. That is, every time the PBR teams were diluted with non-PBR reviewers they tend to perform worse relative to the Checklist technique teams.

Considering data from R1, a permutation test as applied in the original experiment [Basili1996] produces 48.620 distinct ways to assign the reviewers into groups of 9. The group with no dilution had the 15.944th highest test statistic, corresponding to a p-value of 0.33. Concerning R2 data, the group with no dilution had the 6.573th highest test statistic, corresponding to a p-value of 0.14. Therefore, unlike the original study, we cannot reject hypothesis H0 for both replications. For R1+R2 there is an exponential number of distinct assignments of reviewers into groups of 9, preventing such an analysis due to execution time restrictions. Other approaches should be pursued to conduct a similar analysis for multiple experiments.

4.2. OS2' – Do individual PBR or Checklist reviewers find more defects?

When analyzing the data for the individual inspectors, first a statistical analysis using ANOVA for balanced design was performed, followed by a qualitative analysis for replications R1, R2 and R1+R2. The goal of the statistical analysis was to determine whether individual reviewers performed differently when using PBR as compared to Checklist. The dependent variables were individual effectiveness and efficiency. This analysis involved two different factors, or treatments: the Reading Technique (RT) and the Requirement Document (DOC). Three hypotheses were tested with relation to both effectiveness and efficiency.

Group effect or RT X DOC interaction effect

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.

Results in Table 2, for both replications and for the combination R1+R2, show that H0 cannot be rejected for the group effect or for the main effect RT, meaning that there is no statistical evidence that the variables affect effectiveness. Conversely, H0 can be rejected for the main effect DOC, meaning that this variable did influence the results.

Table 2 – ANOVA summary table for individual effectiveness.

Independent Variables	Effectiveness (average percentage MINITAB)			P-value		
	R1	R2	R1+R2	R1	R2	R1+R2
RT X DOC	---	---	---	0.275	0.924	0.353
RT	Checklist=11.417 PBR= 13.346	Checklist=12.050 PBR = 14.294	Checklist=11.733 PBR= 13.820	0.404	0.202	0.144
DOC	ATM = 9.310 PG = 15.453	ATM = 11.412 PG = 14.932	ATM = 10.361 PG = 15.192	0.005✓	0.041✓	0.000✓

Results in Table 3 show that concerning efficiency H0 cannot be rejected for any of the variables in either R1 or R2, i.e., we cannot conclude that the variables affected the results. On the other hand, considering the combination R1+R2, H0 can be rejected for the variable RT, pointing out to the relevance of conducting additional replications and meta-analysis.

Table 3 – ANOVA summary table regarding individual subject efficiency.

Independent Variables	Efficiency (average percentage MINITAB)			P-value		
	R1	R2	R1+R2	R1	R2	R1+R2
RT X DOC	---	---	---	0.417	0.344	0.205
RT	Checklist=2.775 PBR = 3.956	Checklist=3.292 PBR = 3.954	Checklist=3.033 PBR= 3.905	0.101	0.239	0.041✓
DOC	ATM = 2.817 PG = 3.814	ATM = 3.397 PG = 3.849	ATM = 3.107 PG = 3.832	0.131	0.425	0.092

To further study whether Checklist or PBR is in fact more effective and efficient, Table 4 summarizes the data collected concerning defects found (the union of all defects uncovered by individual inspectors) and defect occurrences, as well as average subject effectiveness and efficiency (these metrics are defined in Appendix A).

In the original study, individuals using PBR were significantly more effective for both PG and ATM (efficiency was not addressed in the original study), a result that is partly supported by the results from R1. For R1 and the ATM document, subjects using PBR found a higher percentage of the defects than those using Checklist. This result was not statistically significant at the .05 level (p-value = 0.143), though. For the PG document, subjects using Checklist found a higher percentage of defects, on average, than those using PBR. In this case, the result was also not statistically significant at the .05 level (p-value = 0.911). In terms of efficiency (errors/hour), subjects using PBR were more efficient for both documents, but again this result was not statistically significant at the .05 level (p-value = 0.111 and 0.509 for ATM and PG, respectivelly).

For R2 subjects using PBR found a higher percentage of defects than those using Checklist and both documents, but this result was not statistically significant at the .05 level (p-values = 0.270 and 0.431) respectively. In terms of efficiency, subjects using PBR were more efficient for both documents. This result was not also statistically significant at the .05 level (p-value = 0.137 and 0.875 for ATM and PG, respectivelly).

Similarly, for R1+R2 and both the ATM and PG documents, subjects using PBR found a higher percentage of the defects than subjects using Checklist, but this result was not statistically significant at the .05 level (p-values = 0.063 and 0.659 respectively), i.e., one could not reject the hypothesis that the defects found are not affected by the technique used. In terms of efficiency, subjects using PBR did better on both documents (ATM and PG), but again this was not statistically significant at the .05 level (p-value = 0.026 and 0.544, respectivelly).

Table 4– Comparing results for both requirements documents and both replications.

Document		**ATM**		**PG**	
Technique		**Checklist**	**PBR**	**Checklist**	**PBR**
Metric	**Replication**				
Defects Found/Total Defects	**R1**	15/37 (40.5%)	21/37 (56.8%)	20/32 (62.5%)	14/32 (43.8%)
	R2	17/37 (45.9%)	19/37 (51.4%)	14/32 (46.6%)	20/32 (62.5%)
	R1+R2	22/37 (59.5%)	25/37 (67.6%)	21/32 (65.6%)	23/32 (71.9%)
Occurrences of Deffects/Total Occurrences	**R1**	24/333	38/333	45/288	44/288
	R2	34/333	42/333	40/288	46/288
	R1+R2	58/666	80/666	85/576	90/576
Effectiveness	**R1**	7.21 %	11.41 %	15.63 %	15.28 %
	R2	10.21 %	12.61 %	13.89 %	15.97 %
	R1+R2	8.71 %	12.01 %	14.76 %	15.63 %
Efficiency	**R1**	2.00 %	3.62 %	3.53 %	4.10 %
	R2	2.80 %	3.99 %	3.78 %	3.91 %
	R1+R2	2.40 %	3.81 %	3.66 %	4.00 %

Figure 3 summarizes the effectiveness measures for both replications, which presented similar results for the ATM document, with the PBR technique doing slightly better than Checklist, in agreement with results from previous PBR experiments. It is worth noting that PBR produced slightly better results on the ATM document in R1, while Checklist did slightly better in R2. On the other hand, results from R1 on the PG document conflict with those from the original PBR experiment: as opposed to R2, Checklist did better than PBR. It is also worth noting that results from PBR in the PG document are better in replication R2, contrary to what happened with ATM. In fact, one observes almost an inversion of performances between PBR and Checklist from the first to the second replication for the PG document! Considering both techniques in combination (taking the union of the defects reported), replication R1 produced better results on the ATM document and R2 produced better results on the PG document. As we mentioned previously, considering the requirement documents variable H0 can be rejected at the 0.5 level of significance, meaning that the document did influence the results. Considering both replications in combination, PBR did better than Checklist on both documents.

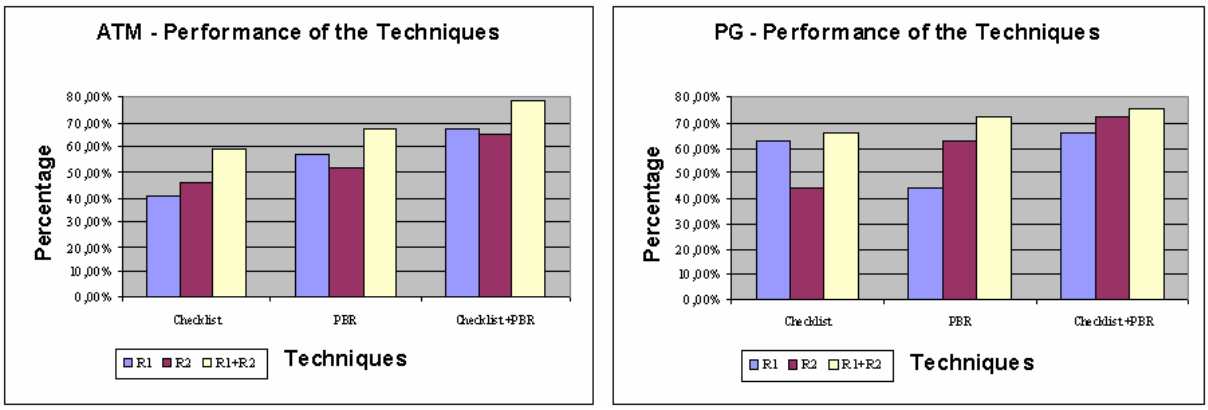

Figure 3 – Performance for both replications.

Figures 4 and 5 present the percentage of defects found by individual subjects in both R1 and R2. Figure 4 shows results for the group that applied Checklist to the ATM document and PBR to the PG document, whereas Figure 5 shows results for the group that applied Checklist to PG and PBR to ATM. A suffix ".1" or ".2" was added to the subject label to identify the replications R1 and R2 which data are in the left and right sides of the graphic, respectively.

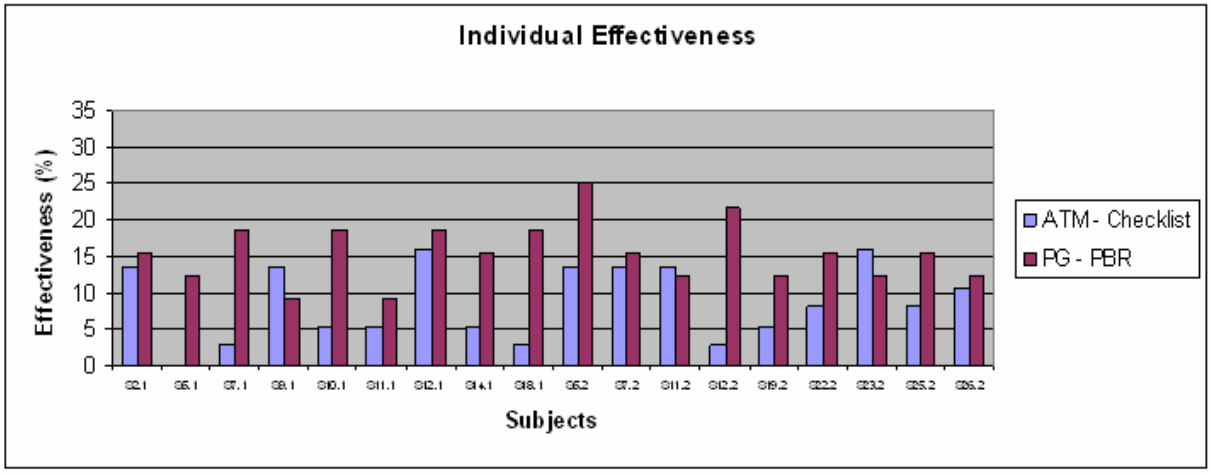

Figure 4 – Subjects individual performance regarding the percentage of defects found: ATM (Checklist) and PG (PBR).

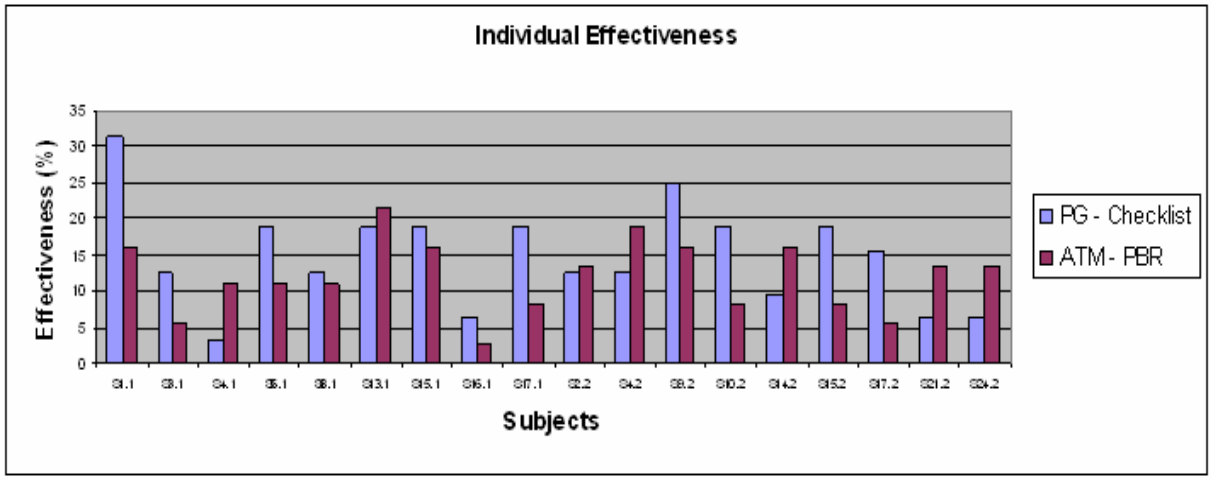

Figure 5 – Subjects individual performance regarding the percentage of defects found: PG (Checklist) and ATM (PBR)

As detailed below, it seems – though such as assumption has no statistical significance –, that there is a direct relationship between subjects' experience and their performance in Checklist. For PBR, such a relationship is not evident at all: the more experienced subjects did not perform as expected, a point that certainly deserves further investigation.

Considering Checklist, in R1, subject S1.1 (Figure 5) achieved the best performance, finding 31.25% of the defects in the PG document. S1.1 is the most experienced subject in the universe of both replications. Considering R2, the most-experienced subject is S10.2 (Figure 5), who did not perform best, identifying less than 20% of the defects in PG. The best performance was by S9.2 (Figure 5), the second-most-experienced subject, who identified 25% of the PG defects. Subject S5.1 from R1 (Figure 4), who has average experience considering the set of subjects, performed worst, finding no defect at all in ATM. Considering R2, subject S12.2 (Figure 4) had performed worst and has one of the lowest experience levels amongst subjects from R2. In R1, in contrast, the least-experienced subject, S15.1 (Figure 5), performed above average, finding almost 20% of the defects in PG.

Considering PBR, subject S5.2 (Figure 4) achieved the best performance identifying 25% of the defects in PG, but has average experience compared to others. The most-

experienced subjects from R1 and R2, S1.1 and S10.2 (Figure 5), found 16% and 8% of the defects in ATM, respectively. Subject S16.1 (Figure 5) performed worst identifying only 2.7% of the total of defects in ATM. S16.1, though still on the average experience level, is more experienced than the subject who performed best, S13.1 (Figure 5).

Figures 6 and 7 present subjects' individual efficiency for both R1 and R2. Figure 6 shows the results for the group that applied Checklist to the ATM document and PBR to the PG document, whereas Figure 7 displays the results for the group that applied Checklist to PG and PBR to ATM.

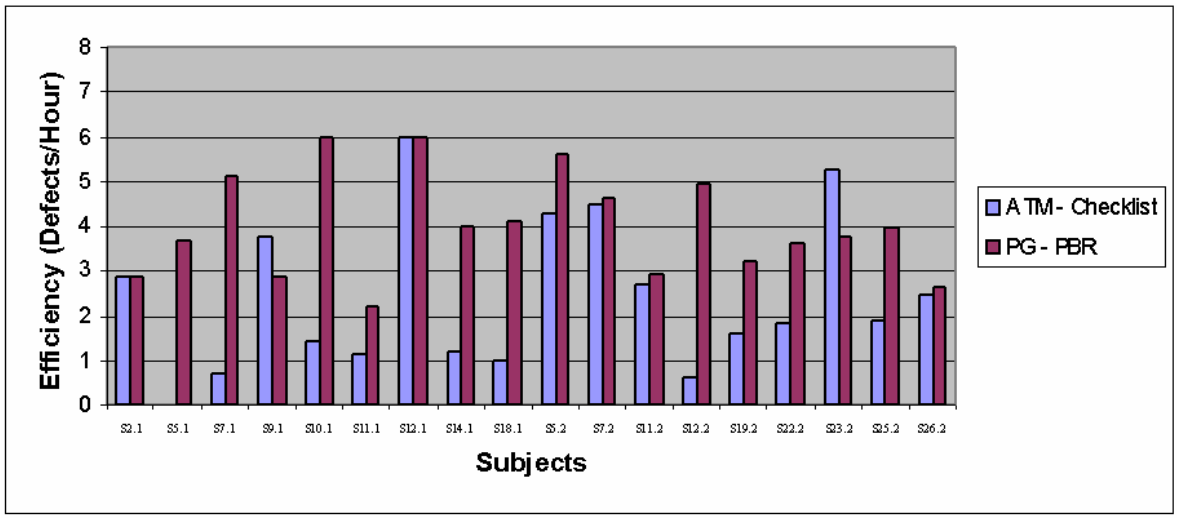

Figure 6 – R1: Individual subject efficiency.

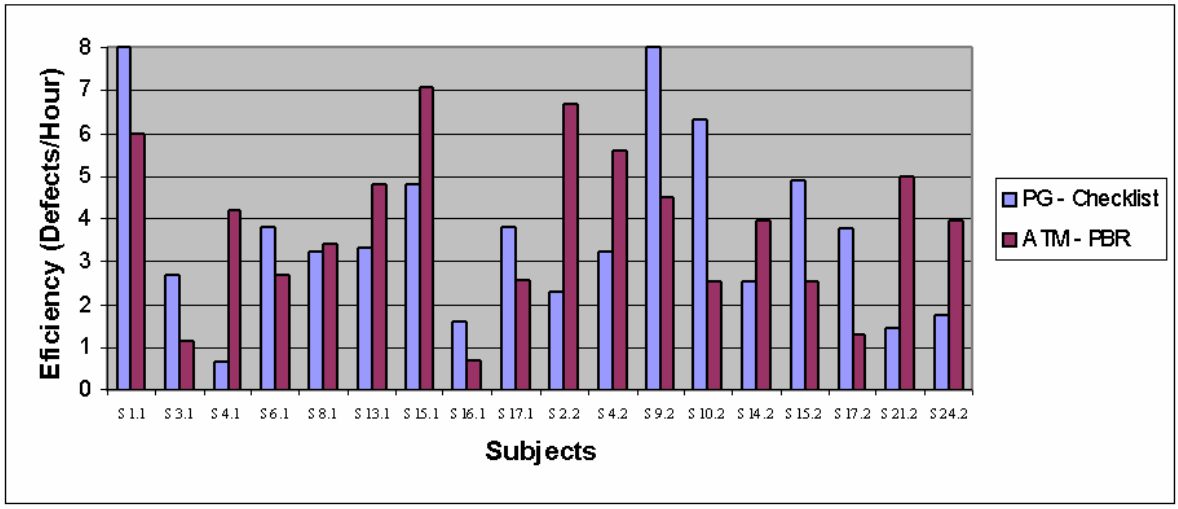

Figure 7 – R2: Individual subject efficiency.

Similar to the performance regarding the number of defects found, it seems that there is a higher relationship between subject's experience and efficiency in Checklist than in PBR, although it is a weak one. Again, for PBR such a relationship is not evident: more experienced subjects did not perform as expected, a point that deserves further investigation.

The best performance with Checklist was by subjects S1.1 in R1, and S9.2 in R2 (Figure 7) who identified 8 defects per hour in PG. Subject S1.1 has the highest experience level in R1, and subject S9.2 has average experience. The most-experienced subject in R2,

S10.2 (Figure 7), found over 6 defects per hour in ATM. As opposite, in R1 S15.1 is the least-experienced (Figure 7) and performed above average, finding almost 5 defects per hour in PG.

Considering PBR, subject S15.1 (Figure 7) performed best, identifying 7.06 defects per hour in ATM. Surprisingly, this is the least-experienced subject, having no previous experience whatsoever. In R1 the most-experienced subject, S1.1 (Figure 7), found 6 defects per hour in ATM, whereas in R2 the most-experienced subject, S10.2 (Figure 7), found less than 3 defects per hour in the same document. Subject S16.1 (Figure 7) performed worst: only 0.67 defects per hour in ATM. Subject S16.1 has no significant previous experience in comparison to the others. Subjects with lower experience levels – S4.2 (Figure 7), S5.2, S11.2, S12.2 and S19.2 (Figure 6) – also performed well in R2: 5.5, 5.5, 3, 5, and 3 defects per hour, respectively.

4.3. OS3' – Does the reviewer's experience affect his or her effectiveness?

Subject's experience in their assigned perspective was measured with a questionnaire, where subjects were asked to indicate the how many years of experience they had on conducting specific tasks related to the three PBR perspectives.

As shown in Figures 8, 9 and 10 for R1, R2 and R1+R2, respectively, there is a weak relationship between experience and effectiveness when using PBR. More experienced reviewers did not perform better, a conclusion supported by the low values obtained when computing the Spearman's and Pearson's correlation tests (under 14%, as shown in Table 5).

Table 5 – Pearson and Spearman's correlation coefficients – PBR effectiveness

Replication	Pearson	Spearman
R1	0.138	0.048
R2	-0.139	0.020
R1+R2	0.036	-0.020

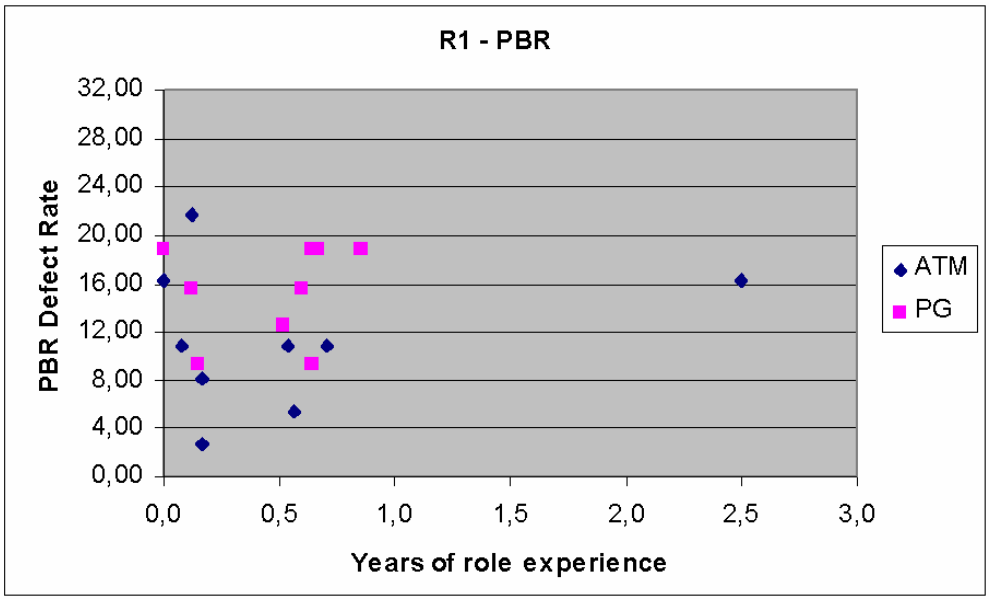

Figure 8 - PBR effectiveness versus readers' role experience (R1)

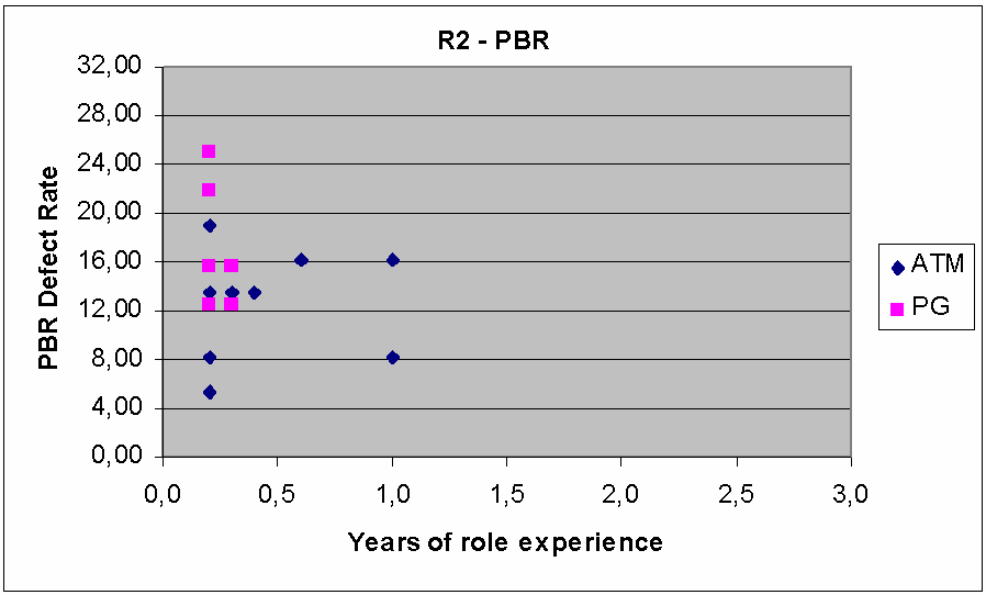

Figure 9 – PBR effectiveness versus readers' role experience (R2)

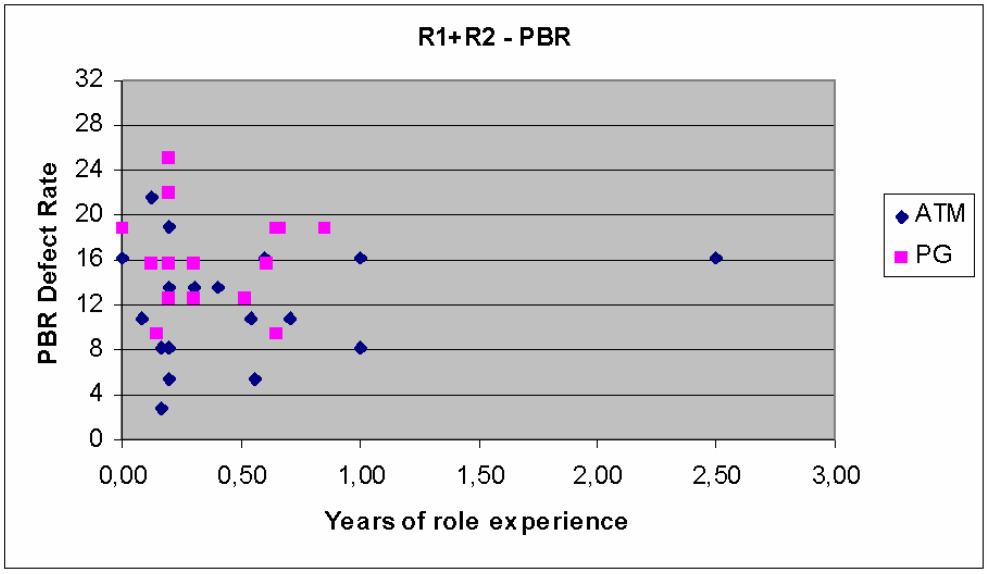

Figure 10 – PBR effectiveness versus readers' role experience (R1+R2)

As shown in Figures 11, 12 and 13 for R1, R2 and R1+R2, respectively, relationship between experience and effectiveness when using Checklist is also weak, and more experienced reviewers did not perform better than the less experienced ones. Again, this is clearly supported by the values of the Spearman's and Pearson's correlation tests (Table 6). Tables 5 and 6 shos that correlation between experience and effectiveness is higher for Checklist than for PBR, though.

Table 6– Pearson and Spearman's correlation coeficients – Checklist effectiveness

Replication	Pearson	Spearman
R1	0.446	0.047
R2	0.319	0.249
R1+R2	0.393	0.057

Figure 11 – Checklist effectiveness versus reader's role experience (R1)

Figure 12 – Checklist effectiveness versus reader's role experience (R2).

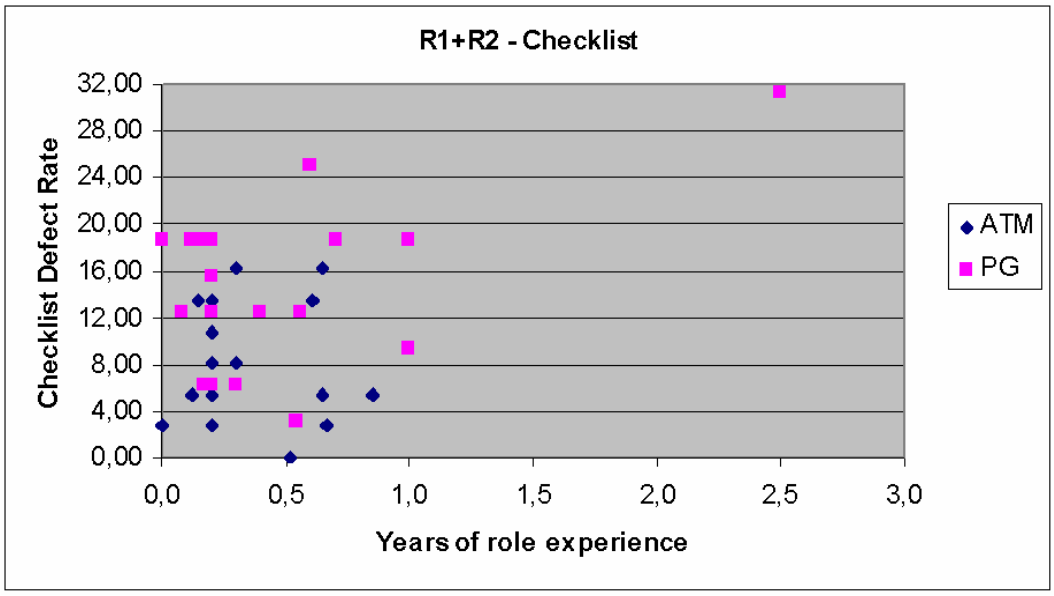

Figure 13 – Checklist effectiveness versus reader's role experience (R1+R2)

4.4. RS1 – Do individual reviewers using PBR and Checklist find different types of defects?

In addition to finding out which technique uncovered more defects, the question of whether users of Checklist and PBR find different defects and types of defects was also addressed. Figure 14 provides an overview of the specific 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 not found by users of the other. Conversely, for the PG document, the techniques appear less complementary: in R1, PBR users found only 1 defect not found by the Checklist users. Overall, considering both techniques together, in R1 subjects found only 25 out of the 37 ATM defects, and 21 out of the 32 PG defects; and in R2 they found only 24 out of the 37 ATM defects and 23 out of the 32 PG defects. Therefore, it may be necessary to complement the Checklist and PBR techniques with other techniques to achieve 100% defect coverage.

Though a similar total number of defects was found in both replications, the sets of defects uncovered in each one are different. Figure 14 shows that a sub-set of seven defects was found in the ATM document by both techniques in both replications. For this particular document and considering Checklist alone, only one defect was found in common in both replications, while PBR alone uncovered three defects in common. In the PG document a sub-set of eight defects was found by both techniques in both replications. For Checklist alone, the replications did not find a single defect in common, while for PBR alone only one common defect was found in both replications. In R1, PBR did not find any defect in the PG document beyond those found by Checklist, but in R2 PBR found eight defects that were not found by Checklist.

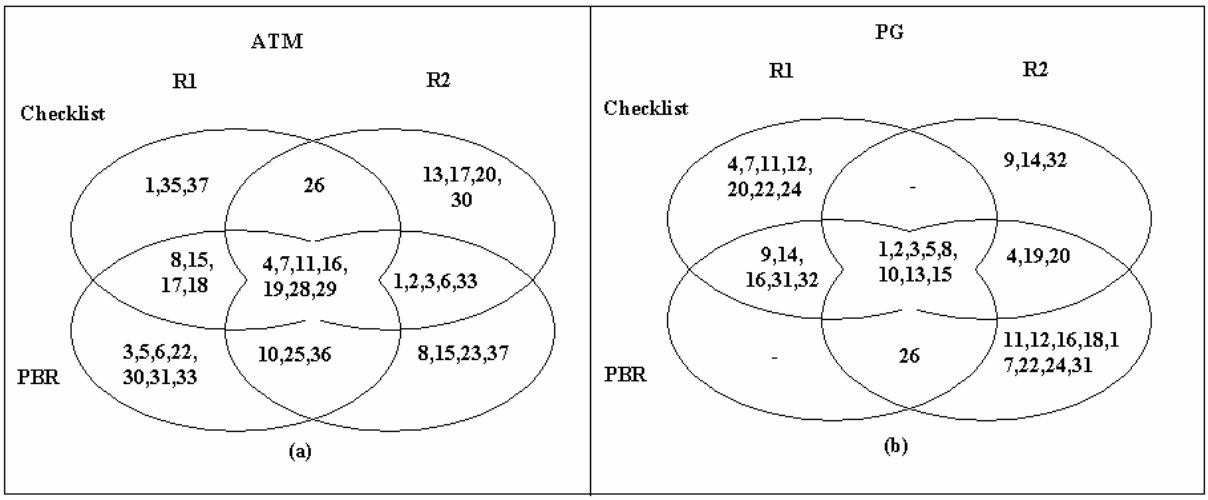

Figure 14 – (a) ATM: defects found per technique; (b) PG: defects found per technique.

Figure 15 provides an overview of the defects identified considering the combination R1+R2. It can be observed that, for both documents, most defects were found by both techniques. This fact does not necessarily imply that the techniques are not complementary, since it could be expected that each defect would be found, even by chance, if we increase the universe of experimentation.

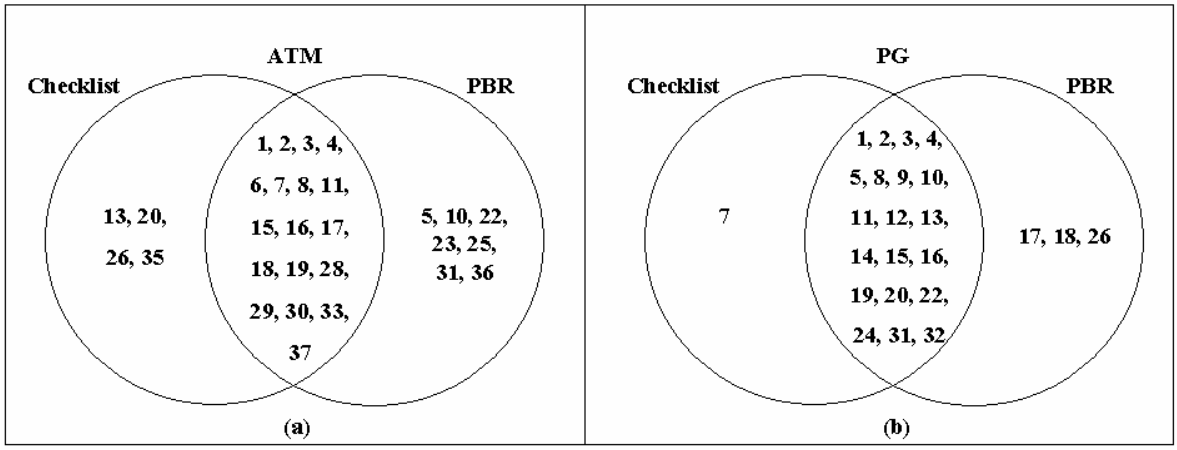

Figure 15 – (a) ATM: defects found per technique; (b) PG: defects found per technique.

We also investigated whether Checklist and PBR users found different *types* of defects. The data shown in Table 7 led to inconclusive results on the adequacy of the techniques for uncovering specific types of defects. For example, for defects of type Ambiguity, subjects using PBR were more effective than those using Checklist in ATM in both R1 and R2, and in the combination R1+R2. However, they were less effective in PG for R1 (Table 11). The suitability of a particular technique to uncover certain defect types 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 the reviewers using that technique identifies defects of a particular type? What about the PBR perspectives?

Table 7 - ATM – Checklist/PBR: Percentage of defects found by defect type.

ATM							
Type	Nº of Defects	% of Defects Found Checklist			% of Defects Found PBR		
		R1	R2	R1+R2	R1	R2	R1+R2
A	8	25.00	37.50	50.00	62.50	50.00	75.00
E	1	100.00	100.00	100.00	0.00	100.00	100.00
II	4	50.00	75.00	75.00	100.00	75.00	100.00
IF	8	62.50	75.00	75.00	75.00	62.50	75.00
MD	2	0.00	0.00	0.00	0.00	50.00	50.00
O	14	35.71	28.57	57.14	42.85	35.71	50.00

Table 8 - ATM – Checklist/PBR: Percentage of defect occurrences by defect type

ATM								
Type	Possible Defect Occurrences		% of Number Occurrences Checklist			% of Number Occurrences PBR		
	R1/R2	R1+R2	R1	R2	R1+R2	R1	R2	R1+R2
A	72	144	2.78	4.17	3.47	9.72	18.08	13.89
E	9	18	11.11	33.33	22.22	0.00	11.11	5.56
II	36	72	19.44	19.44	19.44	16.67	22.22	19.44
IF	72	144	11.11	22.22	16.67	22.22	15.28	18.75
MD	18	36	0.00	0.00	0.00	0.00	5.56	2.78
O	126	252	4.76	3.97	4.37	7.14	6.35	6.75

Tables 7 and 8 present, respectively, the percentage of defects and defect occurrences observed in both replications with the application of Checklist and PBR on the ATM document, organized by defect type. Compared to Checklist, PBR detected more defect occurrences in both replications, except for defects of types Extraneous Information and Incorrect Fact. In terms of defect occurrences PBR did better in R1, except for the types Extraneous Information and Inconsistent Information, while in replication R2 Checklist did better on two types of defects only, Extraneous Information and Incorrect Fact. It is worth noting that the number of occurrences of defects of the Ambiguous and Incorrect Fact types significantly favors PBR in R1. In R2 the number of occurrences of defects of the types Ambiguous and Miscellaneous favors PBR, while the number of occurrences for the types Incorrect Fact and Extraneous Information favors Checklist. Overall, PBR did better on the ATM document.

Table 9 - ATM – PBR: Number of defects found per perspective

ATM – PBR Detailed by Perspective									
Type	Number of Defects								
	Designer			Tester			User		
	R1	R2	R1+R2	R1	R2	R1+R2	R1	R2	R1+R2
A	3	3	4	2	4	5	1	3	4
E	0	1	1	0	0	0	0	0	0
II	3	2	4	1	3	3	1	1	2
IF	3	3	3	4	4	6	4	4	5
MD	0	0	0	0	1	1	0	0	0
O	2	3	4	2	2	3	3	2	5

Table 10 - ATM – PBR: Number of defects occurences per perspective

ATM – PBR Detailed by Perspective									
Type	Number of Occurrences								
	Designer			Tester			User		
	R1	R2	R1+R2	R1	R2	R1+R2	R1	R2	R1+R2
A	4	4	8	2	5	7	1	4	5
E	0	1	1	0	0	0	0	0	0
II	4	2	6	1	4	5	1	2	3
IF	3	3	6	6	4	10	7	4	11
MD	0	0	0	0	1	1	0	0	0
O	3	4	7	3	2	5	3	2	5

Tables 9 and 10 present detailed information on the performance of each PBR perspective on the ATM document, which was quite uniform despite some variations across both replications: i) one occurrence of the only defect of type *Extraneous* and one occurrence of the two defects of type *Miscellaneous* were found only in R2, and ii) in R2 there were almost twice the number of occurrences of defects of type *Ambiguous*, whereas the number of occurrences of type *Incorrect Fact* was higher in R1.

As far as perspectives are concerned, all perspectives did slightly better in R2 and the Tester perspective was the only one that did better or equal in R2 for all defect types concerning the number of defects found.

Table 11 - PG – Checklist/PBR: Percentage of defects found by defect type.

PG							
Type	Nº of Defects	% of Defects Found Checklist			% of Defects Found PBR		
		R1	R2	R1R2	R1	R2	R1R2
A	4	100.00	75.00	100.00	75.00	75.00	100.00
E	1	100.00	0.00	100.00	0.00	100.00	100.00
II	10	60.00	50.00	70.00	30.00	70.00	80.00
IF	2	100.00	100.00	100.00	100.00	100.00	100.00
MD	3	0.00	0.00	0.00	0.00	0.00	0.00
O	12	58.33	33.33	58.33	50.00	58.33	66.67

Table 12 - PG – Checklist/PBR: Percentage of defect occurrences by defect type

PG								
Type	Possible Defect Occurrences		% of Number Occurrences Checklist			% of Number Occurrences PBR		
	R1/R2	R1+R2	R1	R2	R1+R2	R1	R2	R1+R2
A	36	72	16.67	22.22	19.44	13.89	19.44	16.67
E	9	18	22.22	0.00	11.11	0.00	11.11	5.56
II	90	180	13.33	14.44	13.89	4.44	8.89	6.67
IF	18	36	55.56	61.11	58.33	77.78	61.11	69.44
MD	27	54	0.00	0.00	0.00	0.00	0.00	0.00
O	108	216	13.89	7.41	10.65	19.44	17.59	18.52

Tables 11 and 12 present, respectively, the percentage of defects and defect occurrences observed with Checklist and PBR in the PG document, organized by defect type. Compared to Checklist, in R1 PBR did equal or worse in terms of the number of defects

found for all defect types. On the other hand, in R2, PBR did equal or better than Checklist on the number of defects uncovered, for all defect types. In both replications, considering the number of occurrences Checklist did better for defects of types *Ambiguous* and *Inconsistent Information* and PBR did better for defects of types *Incorrect Fact* and *Omission*, performing significantly better for this last type.

Table 13 - PG – PBR: Number of defects found per perspective

PG – PBR Detailed by Perspective									
Type	**Number of Defects**								
	Designer			**Tester**			**User**		
	R1	**R2**	**R1+R2**	**R1**	**R2**	**R1+R2**	**R1**	**R2**	**R1+R2**
A	2	3	4	1	2	2	0	1	1
E	0	0	0	0	0	0	0	1	1
II	0	3	3	3	1	4	1	4	4
IF	2	1	2	2	2	2	2	2	2
MD	0	0	0	0	0	0	0	0	0
O	4	5	7	4	3	4	3	4	5

Table 14 - ATM – PBR: Number of defects occurences per perspective

PG – PBR Detailed by Perspective									
Type	**Number of Occurrences**								
	Designer			**Tester**			**User**		
	R1	**R2**	**R1+R2**	**R1**	**R2**	**R1+R2**	**R1**	**R2**	**R1+R2**
A	4	3	7	1	2	3	0	2	2
E	0	0	0	0	0	0	0	1	1
II	0	3	3	3	1	4	1	4	5
IF	5	3	8	5	5	10	4	3	7
MD	0	0	0	0	0	0	0	0	0
O	6	7	13	8	6	14	7	6	13

Tables 13 and 14 present, respectively, detailed information on each of the three PBR perspectives on the PG document. One observes that the technique performed quite uniformly in both replications. A significant difference favoring replication R2 was observed for the *Inconsistent Information* defect type in terms of the numbers of both defects and defect occurrences. The single defect of type *Extraneous* was found in R2 only.

Regarding the perspectives, the Designer and User perspectives did better in R2, while the Tester perspective did worse. User was the only perspective that did better or equal in R2 in the number of defects, for all defect types.

- Combined Results on the ATM Document:

Figures 16 and 17 show individual subject performance on ATM in both replications, considering Checklist and PBR, respectively, organized by defect type. Data from replications R1 and R2 are on the left and right sides, respectively. It can be observed that for PBR in replication R1 defects of type *Incorrect Fact* were found by most subjects, followed by

defects of type *Omission*. In replication R2 there is a variation and Defects of type *Incorrect Fact* were found by most subjects using Checklist, followed by defects of type *Inconsistent Information*, while most subjects using PBR found more defects of the *Ambiguous* type followed by defects of types *Incorrect Fact* and *Omission*. It is worth noting that these defect types are commonly detected by both techniques.

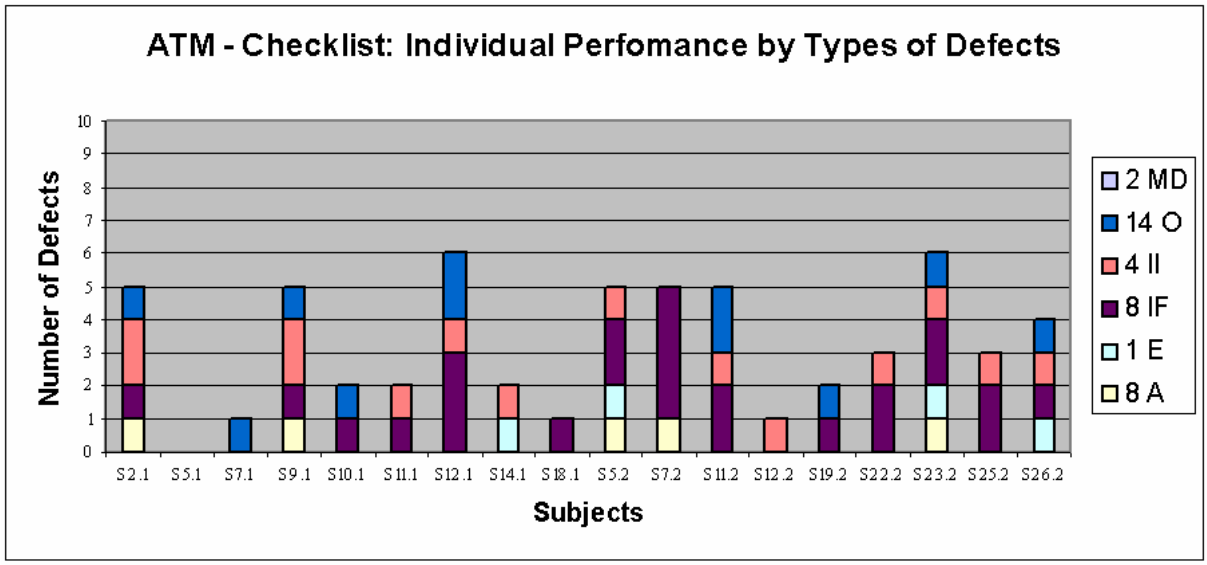

Figure 16 - ATM-Checklist - Types of defects found by each subject.

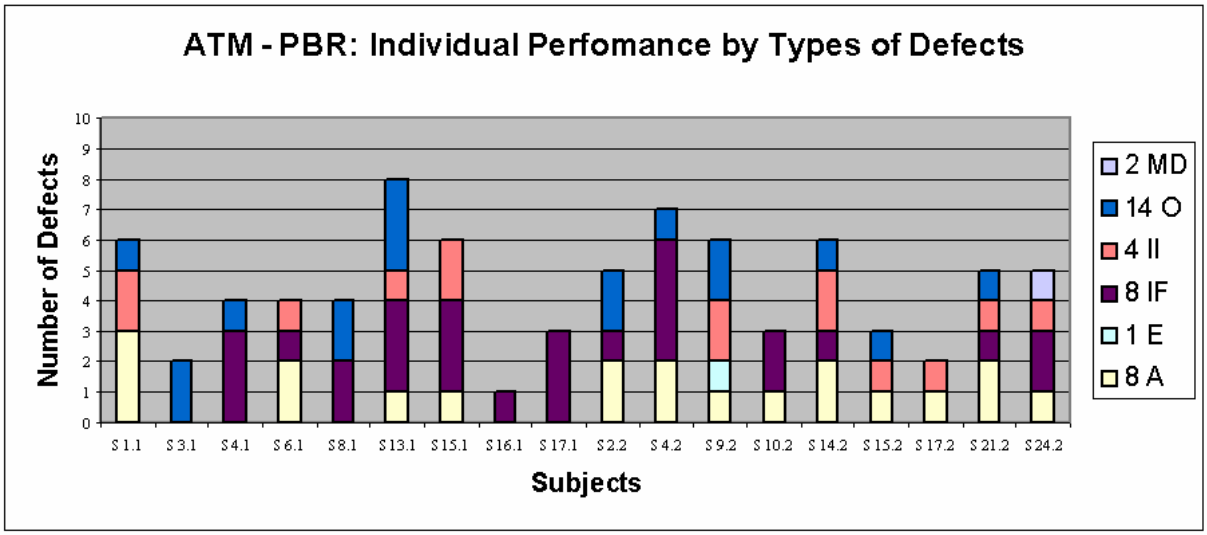

Figure 17 - ATM-PBR –Types of defects found by each subject.

- Combined Results on the PG Document:

Figures 18 and 19 present the individual subject performance in detecting different defect types in the PG document, for both replications, considering Checklist and PBR, respectively. As it can be observed, in R1 defects of type *Incorrect Fact* were found by most subjects using Checklist, followed by defects of the *Inconsistent Information* and *Omission* types. In R2, an equal number of subjects found these three types of defect. Defects of the

type *Miscellaneous* were not found with Checklist. In R1 most subjects using PBR found defects of type *Incorrect Fact*, followed by defects of type *Omission*, whereas in R2 an equal number of subjects found defects of these two types, followed by defects of type *Ambiguous*. Note that defects of the types *Incorrect Fact* and *Omission* were commonly found by both techniques, and that *Incorrect Fact* is again one of the two defect types most commonly identified in both replications.

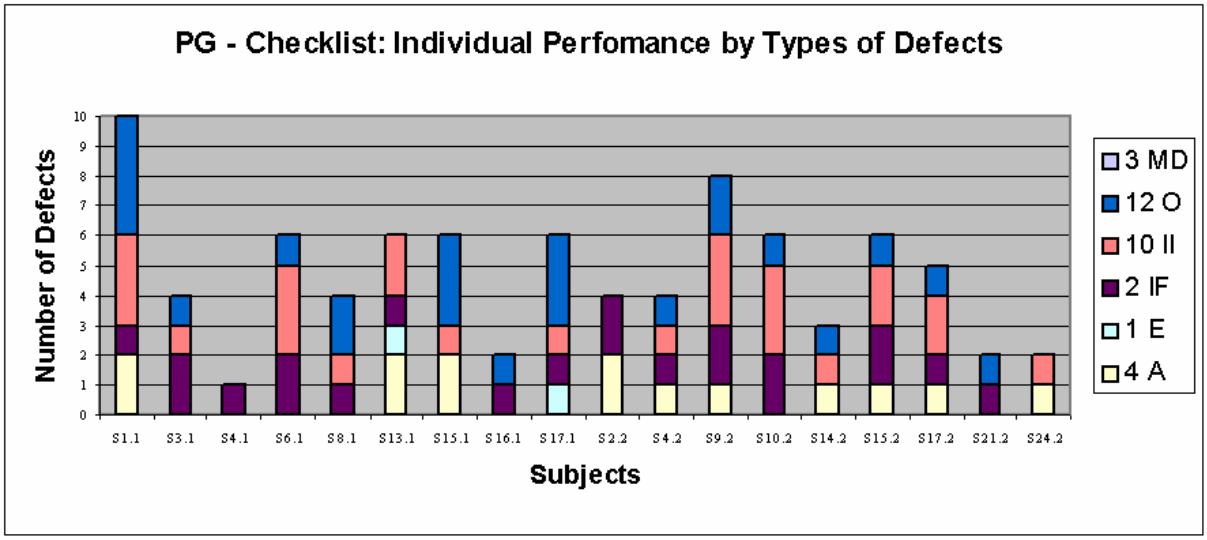

Figure 18 - PG-Checklist - Types of defects found by each subject.

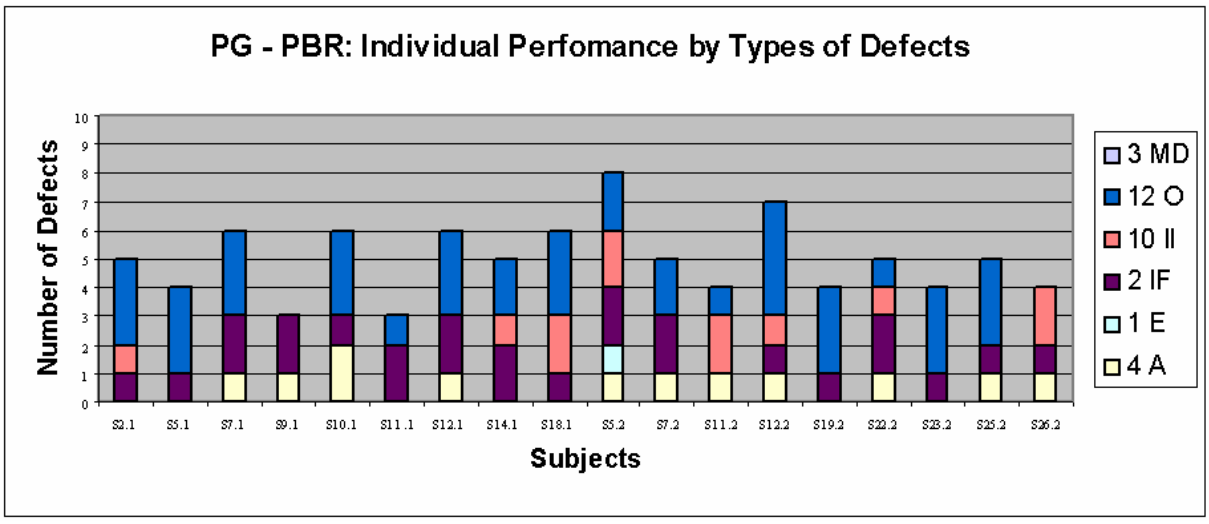

Figure 19 - PG-PBR - Types of defects found by each subject.

In summary, it seems that both techniques are equally suitable for identifying defects of the types *Incorrect Fact* and *Omission*. Analysis of the previous figures raises some intriguing questions, such as: “Is there a type of defect for which a technique would be more adequate than another?”; “Is the technique uniform in the sense that it would lead the majority of subjects identifying a particular type of defect?”. Although the data available is insufficient to drive definite conclusions, it clearly provides a contribution in this direction.

4.5. RS2 – Do the PBR perspectives have the same effectiveness?

Another open question concerned the effectiveness of the three PBR perspectives. Table 15 summarizes the effectiveness (percentage of defects found) and efficiency (defects per hour) for the reviewers of R1 using each PBR perspective for both the ATM and PG documents. Reviewers using the Designer perspective performed better on the ATM document considering both effectiveness and efficiency. On the PG document reviewers using the Tester perspective were the most effective while reviewers using the Designer perspective were the most efficient. Data from both documents was combined and an ANOVA was run to test whether the perspective had a significant effect on either effectiveness or efficiency. Results showed no significant influence on either the effectiveness or efficiency ($p = 0.654$, $p = 0.128$). From this data one cannot draw any conclusions about the comparative effectiveness of the perspectives.

Table 15 – ATM/PG – PBR: R1 Average percentage of defects found and defect observation rate.

Perspective	Percentage of defects found (average)			Defect-observation rate (defects/hour)		
	ATM	PG	ATM+PG	ATM	PG	ATM+PG
Designer	**12.61**	15.63	14.11	**4.73**	**4.95**	**4.84**
Tester	10.81	**17.71**	**14.26**	3.44	4.41	3.89
User	10.81	12.50	11.65	2.68	2.92	2.80
Average	**11.41**	**15.28**	**13.34**	**3.62**	**4.10**	**3.86**

Table 16 summarizes the effectiveness (percentage of defects found) and efficiency (defects per hour) for the reviewers of R2 using each PBR perspective for ATM and PG. Reviewers using the Tester perspective were the most effective on the ATM document and reviewers using the Designer perspective were the most efficient. On the PG document reviewers using both the Designer and User perspectives were the most effective while reviewers using the Tester perspective were the most efficient. Again, data from both documents was combined and an ANOVA was run, with the results showing no significant influence of the perspective on either effectiveness or efficiency ($p = 0.945$, $p = 0.642$). From this data one cannot draw any conclusions on the comparative effectiveness of the perspectives.

Table 16 – ATM/PG – PBR: R2 Average percentage of defects found and defect observation rate.

Perspective	Percentage of defects found (average)			Defect-observation rate (defects/hour)		
	ATM	PG	ATM+PG	ATM	PG	ATM+PG
Designer	12.61	**16.67**	14.64	**4.56**	3.84	**4.20**
Tester	**14.41**	14.58	**15.50**	4.30	**3.99**	4.15
User	10.81	**16.67**	13.74	3.13	3.90	3.52
Average	**12.61**	**15.97**	**14.63**	**3.99**	**3.91**	**3.95**

Table 17 summarizes effectiveness (percentage of defects found) and efficiency (defects per hour) of the reviewers using each PBR perspective on ATM and PG, for R1+R2. For both documents, reviewers using the Tester and the Designer perspectives had the same performance considering the percentage of defects found, and the Designer perspective was the most efficient. After combining the data from both documents running an ANOVA results showed no significant influence of the perspective on either effectiveness or efficiency ($p = 0.641$, $p = 0.086$). Thus, one cannot draw any conclusions on the comparative effectiveness of the perspectives from this data.

Table 17 – ATM/PG – PBR: R1+R2 Average percentage of defects found and defect observation rate.

Perspective	Percentage of defects found (average)			Defect-observation rate (defects/hour)		
	ATM	PG	ATM+PG	ATM	PG	ATM+PG
Designer	**12.61**	**16.15**	**14.38**	**4.64**	**4.40**	**4.52**
Tester	**12.61**	**16.15**	**14.38**	3.87	4.20	4.04
User	10.81	14.58	12.70	2.90	3.41	3.16
Average	**12.01**	**15.63**	**13.82**	**3.81**	**4.00**	**3.91**

4.6. RS3 – Do the PBR perspectives find different types of defects?

Finally, we addressed the question of whether the sets of defects found by the perspectives were orthogonal. In other words, do the perspectives complement each other, or do they all tend to find the same defects? If perspectives are complementary, then there is a benefit from using the entire collection, although using multiple reviewers is more expensive. Figures 20 and 21 show the data for R1 and each Requirements Document:

- Part (a) shows, for each perspective, which defects were found by the perspective and the number of defect occurrences found (in parenthesis), e.g., in Figure 20, defect 3 was found by at least one Designer and at least one Tester but by no Users, and the Designers found 11 different defects and 14 defect occurrences;
- Part (b) shows which perspective(s) found the greatest number of occurrences of each defect, e.g. in Figure 20, defect 3 was reported more times by Designers than by Testers.

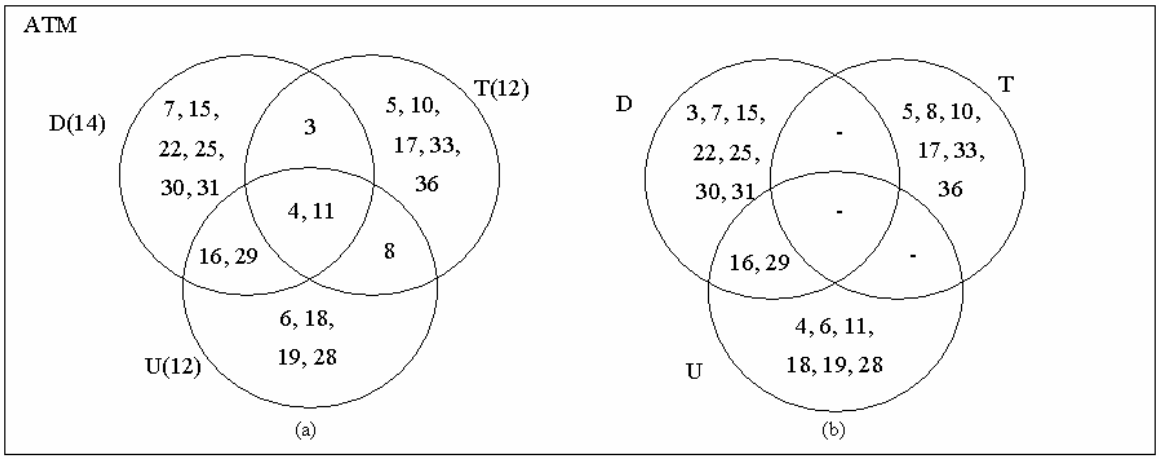

Figure 20 – R1/ATM (a) Identification of the different defects found by the perspectives; (b) Grouping the defects by perspective that obtained better or equal performance than the other perspectives.

Figure 20 (a) shows that a sub-set of two defects was found by all three perspectives. The Designer perspective identified 11 out of 37 defects (29.73%), while the Tester and the User perspectives identified 9 out of 37 (24,32%). Figure 20 (b) shows that the Designer perspective also performed better than the others for seven defects, and performed as well as the User perspective for two defects.

The Designer perspective identified the greater number of defect occurrences, 14, while the Tester and the User perspectives identified 12, as shown in Figure 20 (a), summing up 38 (=14+12+12) out of 333 possible defect occurrences identified. For replication R1 and the ATM document, the perspectives appear to be complementary.

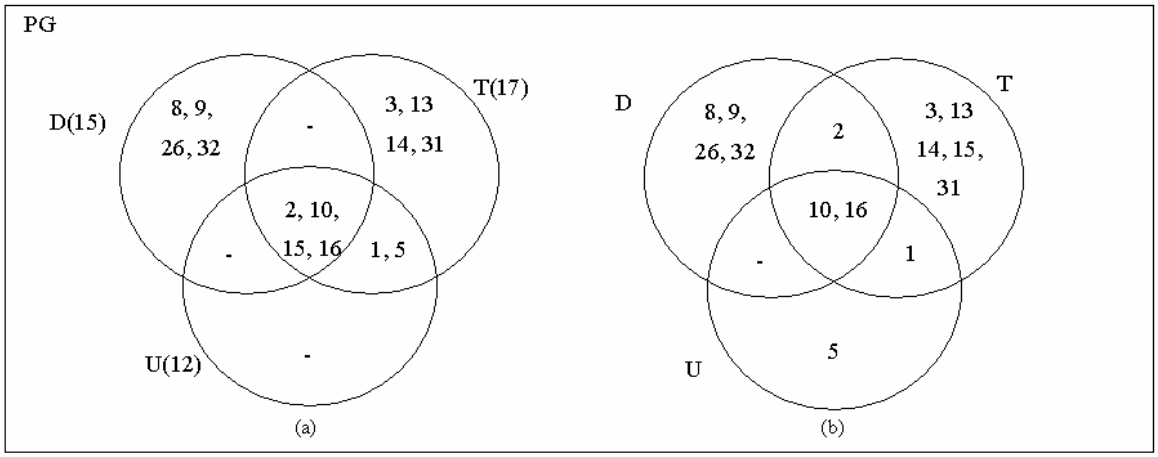

Figure 21 – R1/PG (a) Identification of the different defects found by the perspectives; (b) Grouping the defects by perspective that obtained better or equal performance than the other perspectives.

In Figure 21 the same previous views of Figure 20 are depicted for PG. All three perspectives found four specific defects in common. The Tester perspective identified 10 (31.25%) out of 32 defects, while the User and the Designer perspectives identified 6 and 8 defects, respectively. From Figure 21 (b) it can be observed that the Tester perspective also performed better than the other ones for 5 defects, while for two of the defects all perspectives performed equally. From Figure 21 (a) it can also be observed that the Tester perspective identified more defect occurrences, 17, while the Designer performed better than the User, identifying 15 defect occurrences against 12 by the User. These sum up to 44 defect occurrences detected out of 288 possibilities. For the PG document, the Designer and Tester perspectives appear complementary, but the User perspective does not add much benefit.

Figure 22 shows the equivalent views for R2. Figure 22 (a) shows that six defects were found by all three perspectives. The Tester perspective identified 14 out of 37 defects (37.84%), while the Designer and the User perspectives identified 12 and 10 defects, respectively. The Tester perspective also identified more defect occurrences, 16 out of 333, while the Designer and the User identified 14 and 12, respectively, as shown in Figure 22 (a). This figure also shows that a total of 42 (=16+14+12) defect occurrences were detected out of 333 possible occurrences. Figure 22 (b) shows that the best performance was by the Tester perspective for four defects, and also that all perspectives performed equally well for four of the defects.

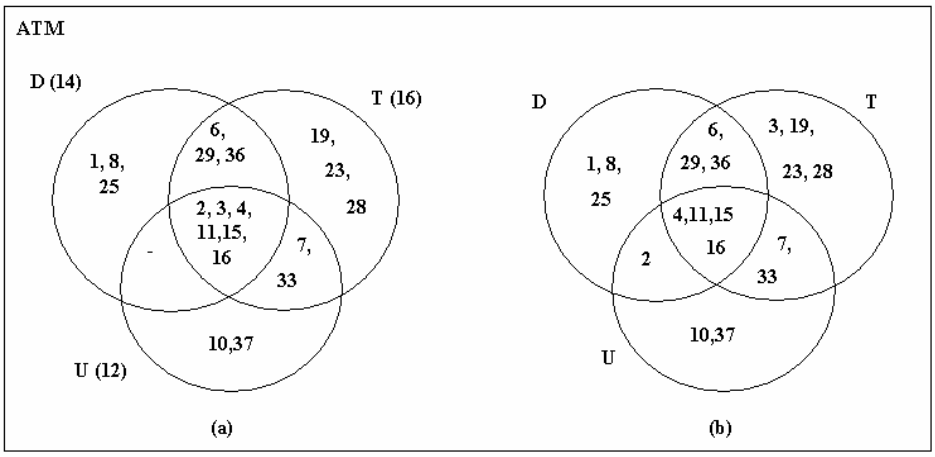

Figure 22 – R2/ATM (a) Identification of the different defects found by the perspectives; (b) Grouping the defects by perspective that obtained better or equal performance than the other perspectives.

Figure 23 shows the data for R2 and the PG document. All three perspectives found four specific defects in common. The User and Designer perspectives identified 12 defects out of 32 (37.5%), while the Tester perspective identified 8. From Figure 23 (b) it can be observed that the User perspective performed best for seven defects, and that all perspectives performed equally well for a single defect. Both the User and Designer perspectives identified 16 defect occurrences, while the Tester identified 14. Thus, a total of 46 defect occurrences were detected out of 288 possibilities.

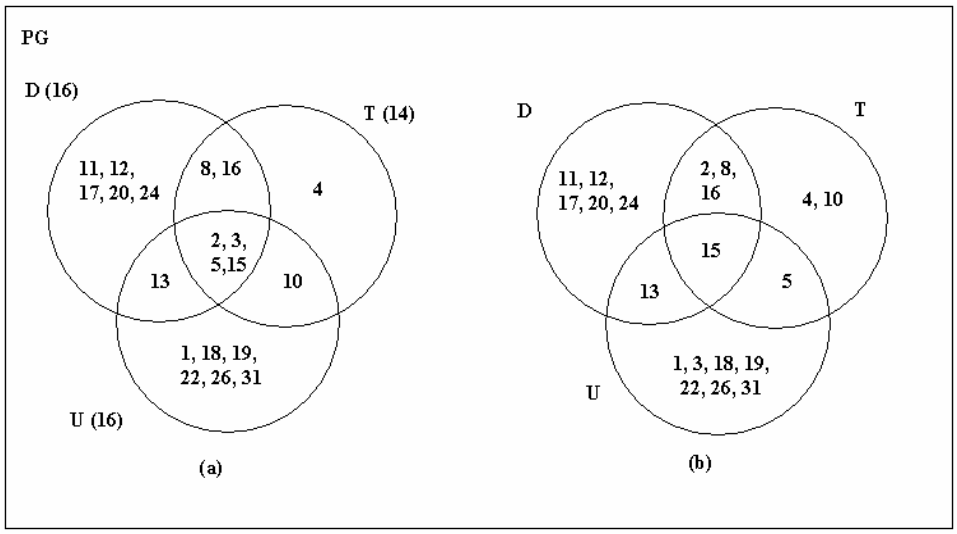

Figure 23 – R2/PG (a) Identification of the different defects found by the perspectives; (b) Grouping the defects by perspective that obtained better or equal performance than the other perspectives.

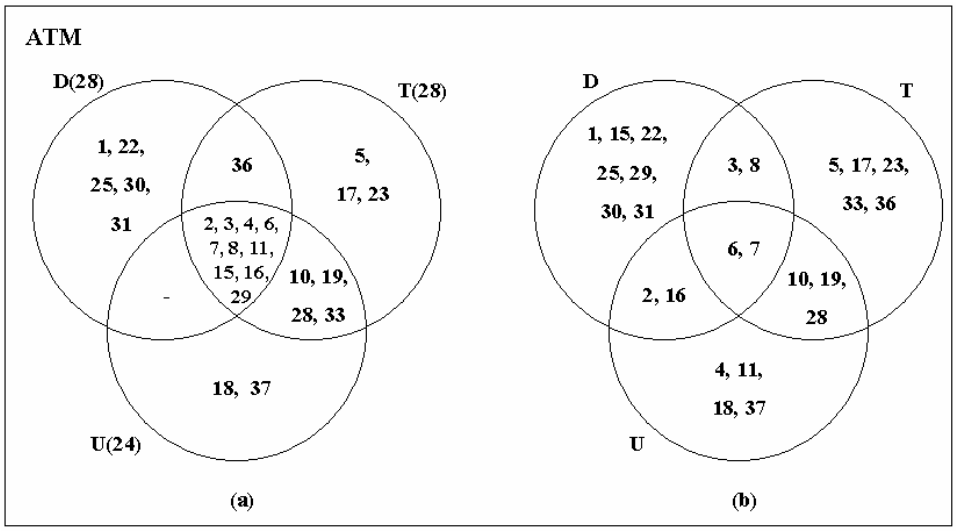

Figure 24 – R1+R2/ATM (a) Identification of the different defects found by the perspectives; (b) Grouping the defects by perspective that obtained better or equal performance than the other perspectives.

In Figure 24 (a) we notice that all three perspectives found ten specific defects in common. The Tester perspective identified 18 (48.65%) out of 37 defects, while both perspectives, User and Designer identified 16 defects. From Figure 24 (b) it can be observed that the Designer perspective performed better than the other ones for 7 defects, while the Tester and the User perspectives performed better for 5 and 4 defects, respectively. From Figure 24 (a) we can also observe that the Designer and Tester perspectives identified more defect occurrences, 28, while the User identified 24. These sum up to 80 defect occurrences detected out of 666 possibilities.

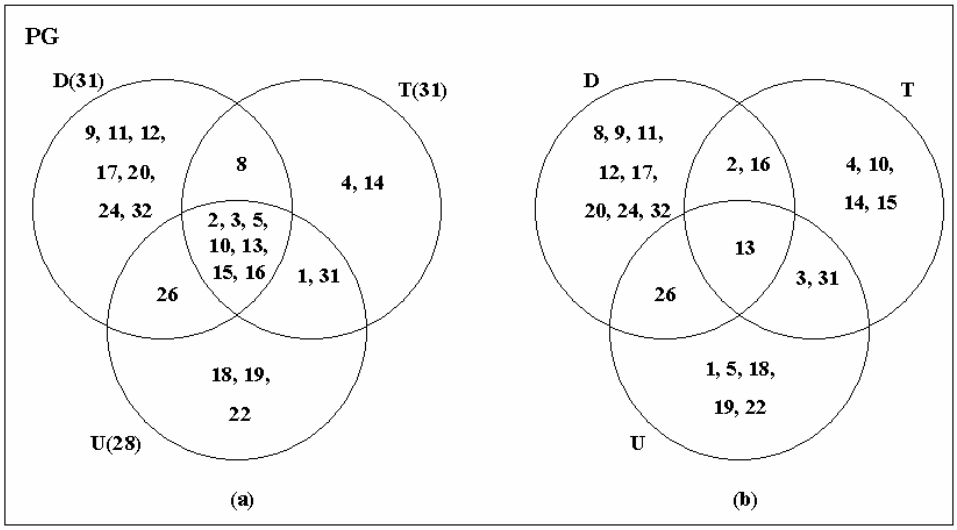

Figure 25 – R1+R2/PG (a) Identification of the different defects found by the perspectives; (b) Grouping the defects by perspective that obtained better or equal performance than the other perspectives.

In Figure 25 we can observe that all three perspectives found seven specific defects in common. The Designer perspective identified 16 (50.00%) out of 32 defects, while the User

identified 13 and the Tester perspective identified 12. From Figure 25 (b) it can be observed that the Designer perspective also performed better than the other ones for 8 defects, while User and Tester identified 5 and 4, respectively. From Figure 25 (a) it can also be observed that the Designer and Tester perspectives identified more defect occurrences, 31, while the User identified 28 defect occurrences. These sum up to 90 defect occurrences detected out of 576 possibilities.

4.7. Feedback from Subjects

After the replication all subjects received a Feedback Questionnaire to evaluate different aspects of the experiment: training time and quality, domain knowledge, form adequacy, etc. The more frequent and relevant comments are presented in Table 18.

Table 18 – Subjects comments about the experiment.

Comments	Number of Subjects	
	R1	R2
Enough time was allocated for training and execution.	15	10
PBR was easier to apply because it is more specific, as it defines a perspective/role for each reader.	10	1
The ATM document was easier to handle because it describes a familiar application domain, even though it has more detailed functional requirements.	10	4
The PG document was difficult to work with because automated parking garages systems are an unfamiliar domain, although it has simpler functional requirements[1].	7	7
It was not difficult to understand the requirements documents.	3	1

5. Threats to Validity

In this section we raise and discuss issues that may represent threats to the validity of the results discussed in this text. Two specific actions, already mentioned in Section 3.1, were taken to minimize possible threats derived from the fact that these experiments were the first ones carried out by the Brazilian replicators: i) close interaction with the original experimenters; and ii) the conduction of a pilot study to master how to run the experiment and to debug its techniques and procedures. Others issues can be considered as internal and external threats to the validity of these replications:

- Internal threats:
  - The native language of the subjects was Portuguese, whereas class lecture notes, assignment instructions, techniques and artifacts were all in English. Therefore, lack of English proficiency might affect the results of the study.
  - The replicators did two relevant changes to the experimental procedures before running the studies: firstly, they adjusted the training time, but keeping it equal for both techniques regardless of the fact that technique complexity is not equivalent; they differ in the level of detail and of required background knowledge. Secondly, in

[1] There are not many parking garages like the one described in the experiment in Brazil, which explains why many subjects considered the PG document more difficult. This is an important aspect to consider when investigating the role of different cultural settings in experiment replication.

R1 the training and the application of the techniques happened in the same day, whereas in R2 they happened in two days.

- Although the role of the trainer has not been explicitly addressed in this text, the trainer expertise and experience is another factor of replication process conformance and may impact the results. Brazilian trainers were neither as experienced in empirical studies as the original experimenters nor were they as experts in the PBR technique as the USA partners.

- External threats:
  - The replications were run in university classrooms and subjects were students, obviously not as experienced as industrials professionals. Moreover, most subjects were inexperienced in their PBR perspective, which does not permit a direct transfer of the conclusions to industrial inspectors.
  - The native language, as mentioned before, as well as application domain knowledge may also constitute issues that should be further addressed in future experiments.

6. Summary, Insights and Further Work

Results from these replications are in partial agreement with previous results by Basili [Basili1996]. Specifically, PBR reviews proved more efficient for both documents in both replications; PBR reviews were more effective for the ATM document and as effective as Checklist for the PG document in replication R1; and were more effective for both documents in replication R2. Although apparently pointing to potential benefits of PBR, these experiments raise a number of issues for further investigation. There is an obvious conflict between results from both replications for the PG document, and there is also an indication of the complementary nature of both inspection approaches, since each one uncovered defects that the other did not.

The uniformity of the results obtained with each approach is a major concern. A crucial goal for software process improvement is to make software development results more repeatable and less dependent on characteristics of individual developers. Neither Checklist nor PBR led to complete uniformity of defect reporting, but with PBR a higher percentage of subjects achieved the same higher performances (within each perspective) in both replications.

Simple and relevant questions not addressed in this text arise from these replications: Have subjects mastered the technique? Are they really applying it? Which are their real background, experience, and abilities? Answering such questions requires running more carefully-designed and well-planned experiments addressing key issues related to subject characterization, technique application conformance and supporting meta-analysis.

Subject characterization based on direct subject survey poses a threat to the validity of the analysis, because there is a variation on the way people perceive and fill these forms. An approach to mitigate this problem is to adopt measures to characterize subjects. A set of well-established and available measures might be defined, such as number of courses taken, grades obtained, etc. Another, maybe better possibility, is to formally pre-test subjects' skills on the techniques under study. Finding out how well subjects have mastered the techniques is another important issue, which might be achieved by creating standard assessing tests.

We believe that technique-application conformance may be improved by properly motivating subjects on the significance of the experiments for constructing a body of knowledge on the technique. Subjects must be aware that, if a real contribution is to be made, taking part in the experiments implies finding defects using the assigned technique, rather than just finding defects. Technique-application conformance must also be better assessed. The feedback questionnaire can be improved for this purpose and interview sessions should always bring out, if not focus on, this particular issue. Mechanisms to track, or even enforce, technique-application conformance should be considered.

Data collection must be improved to support meta-analysis. For example, our experiments suggest that the perspectives are complementary, contradicting previous results by other researchers [Regnell2000]. Additional experiments should be conducted to analyze this issue, but they should be carefully planned to support this type of meta-analysis. Better methodologies for conducting meta-analysis are also necessary. We are currently experimenting with visual data exploration tools to support the analysis of results [Mendonça1999]. Visual approaches complement traditional statistical analysis, bringing the possibility of better exploring the many intervening factors that can significantly affect the results of such experiments.

The PBR Laboratory Package, particularly its training material, evolved as a consequence of the first two replications, and was applied in two other replications, one involving graduate students in an academic setting, another involving professional software engineers in the Telecom Industry.

It is our belief that web-based environments should be used to support experiments. Such an environment can assist the realization of large and consistent multi-institutional experiments that can grant the research community with access to large sets of consistent experimental data. Cooperation networks such as the International Software Engineering Research Network, integrated by independent researchers sharing common interests, may play an essential role on the realization of such large-scale experiments. Moreover, such environments may also contribute to achieve replication process conformity, an issue not explicitly addressed in most the replication studies so far.

Acknowledgements
We would like to thank the Readers' Project team for valuable comments that considerably improved this paper, and the subjects who participated in the studies. This work was conducted under the financial support of CNPq and NSF grants CCR9900307 and CCR0086078 (establishing the Center for Empirically Based Software Engineering).

References

[Basili1987]	Basili, V. and Selby, R. Comparing the Effectiveness of Software Testing Strategies. IEEE Transactions on Software Engineering, 13(12): 1278-1296, 1987.
[Basili1996]	Basili, V.; Green, S.; Laitenberger, O.; Lanubile, F.; Shull, F.; Sorumgard, S. and Zelkowitz, M. The empirical investigation of perspective-based reading. Empirical Software Engineering: An International Journal, 1(2): 1996. pp. 133-164, 1996.
[Basili2001a]	Basili, V.R.; Lindvall, M. and Costa, P. Implementing the Experience Factory concepts as a set of Experience Bases. Proc. 13th International Conference on Software Engineering & Knowledge Engineering, 102-109, 2001.
[Basili2001b]	Basili, V.R.; Tesoriero, R.; Costa, P.; Lindvall, M.; Rus I.; Shull, F. and Zelkowitz, M.V. Building an Experience Base for Software Engineering: A report on the first CeBASE Workshop. Profes (Product Focused Software Process Improvement), 110-125, 2001.
[Box1978]	Box, G. E. P.; Hunter, W. G.; Hunter, J. S. Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building. John Wiley & Sons, 1978.
[Doria2001]	Dória, E. S. Software Engineering Experiment Replications. ICMC-USP, São Carlos, Brazil. Master Thesis, 2001. (In Portuguese)
[Fusaro1997]	Fusaro, P.; Lanubile, F. and Visaggio, G. A replicated experiment to assess requirements inspections techniques. Empirical Software Engineering: An International Journal, 2(1): 39-57, 1997.
[Goth2001]	Goth, G. New Center will help Software Development "Grow Up". IEEE Software, May/June, 2001.
[Laitenberger2000a]	Laitenberger, O.; Atkinson, C.; Schlich, M. and El Emam, K. An experimental comparison of reading techniques for defect detection in UML design documents. Journal of System and Software, 53: 183-204, 2000.
[Laitenberger2000b]	Laitenberger, O.; El Emam, K. and Harbich, T. An Internally Replicated Quasi-Experimental Comparison of Checklist and Perspective-based Reading of Code Documents. IEEE Transactions on Software Engineering, 2000.
[Lott1997]	Lott, C. and Rombach, D. A Repeatable Software Engineering Experiment for Comparing Defect-Detection Techniques. Journal Empirical Software Engineering (1)3, 1997.
[Maldonado2001]	Maldonado, J.C.; Martiniano, L. A. F.; Dória, E.S.; Fabbri, S.; Mendonça Neto, M. Readers Project: Replication of Experiments –A Case Study Using Requeriments Documents. In: ProTeM-CC-Project Evaluation Workshop – International Cooperation, CNPq, Rio de Janeiro, RJ, October, pp. 85-117, 2001.
[Mendonça1999]	Mendonça Neto, M. G. and Sunderhaft, N. L. A State of the Art Report: Mining Software Engineering Data. Rome, NY: U.S. Department of Defense (DoD) Data & Analysis Center for Software, 1999, also available at http://www.dacs.dtic.mil/techs/datamining/.
[Minitab2000]	Minitab Corporation. http://www.minitab.com. Available on-line. Last access on 23-01-2003.
[Porter1995]	Porter, A.; Votta, L. and Basili, V. Comparing Detection Methods for Software Requirements Inspections: A Replicated Experiment. IEEE

	Transactions on Software Engineering 21(6): 563-575, 1995.
[Regnell2000]	Regnell, B.; Runeson, P. and Thelin, T. Are the Perspectives Really Different? – Further Experimentation on Scenario-Based Reading of Requiments. Empirical Software Engineering: An International Journal, 5(4): 2000. pp. 331-356, 2000.
[Shull2002]	Shull, F.; Basili, V.; Carver J.; Maldonado, J.C.; Travassos, G.H.; Mendonça, M. and Fabbri, S. Replicating Software Engineering Experiments: Addressing the Tacit Knowledge Problem. 2002 International Symposium on Empirical Software Engineering (ISESE'02) October 03 - 04, Nara, Japan, 2002.
[Shull2000]	Shull, F.; Rus, I. and Basili, V. B. How Perspective-Based Reading can improve requirements inspections. IEEE Computer 33(7), 2000.
[Shull2001]	Shull, F.; Carver, J. and Travassos, G. H. An Empirical Methodology for Introducing Software Processes. Proc. European Software Engineering Conference, Vienna, Austria,, 288-296, 2001.
[Winer1991]	Winer. B. J., Brown, D. R, and Michels, K. M. Statistical Principles in Experiment Design. New York, NY, 3rd Edition, McGraw-Hill Inc, 1991.
[Zhang1999]	Zhang, Z.; Basili, V. and Shneiderman, B. Perspective-based Usability Inspection: An Empirical Validation of Efficacy. Empirical Software Engineering. An International Journal 4(1): 43-70, 1999.

Annex A – Experimental Design

The experimental design of the PBR experiment is shown in Figure A1. Subjects were divided into two groups of nine people. Both groups applied Checklist on the first day and PBR on the second. The order of utilization of the experimental artifacts – the Parking Garage and Automated Teller Machine requirements – was switched between the two groups. Before applying the techniques on those artifacts, subjects were trained using the ABC Video System requirements document. The groups applying PBR were divided into three subgroups of three subjects. Each subgroup applied the technique from one of the perspectives, either a Designer, a Tester or a User.

<table>
<tr><td></td><td colspan="3">Group 1 – 9 Subjects</td><td colspan="3">Group 2 – 9 Subjects</td><td></td></tr>
<tr><td rowspan="3">First Day</td><td colspan="6">Theory Checklist</td><td rowspan="3">Checklist</td></tr>
<tr><td colspan="3">Training (ABC video)</td><td colspan="3">Training (ABC video)</td></tr>
<tr><td colspan="3">ATM</td><td colspan="3">PG</td></tr>
<tr><td rowspan="4">Second Day</td><td>Designer
3 Subjects</td><td>Tester
3 Subjects</td><td>User
3 Subjects</td><td>Designer
3 Subjects</td><td>Tester
3 Subjects</td><td>User
3 Subjects</td><td rowspan="4">PBR Technique</td></tr>
<tr><td colspan="6">Theory PBR</td></tr>
<tr><td colspan="3">Training (ABC video)</td><td colspan="3">Training (ABC video)</td></tr>
<tr><td colspan="3">PG</td><td colspan="3">ATM</td></tr>
</table>

Figure A1 - Experimental Design.

Four metrics were used to evaluate the data collected:

- **Defects Found**: the number of defects found using a specific technique;
- **Occurrences of Defects**: how many times the defects were observed. The maximal number of occurrences is determined by the number of defects in a given requirement document multiplied by the number of subjects. The total number of defect occurrence (TotalOc) is calculated as following:

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

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

Obviously, the number of defects and occurrences determined by a given subject are the same. This measure evaluates the uniformity of reviewers' results using the same technique or perspective. In the best scenario, all the subjects would

uncover all defects, so we would have the maximal number of occurrences. On the other hand, if different reviewers find distinctly different subsets of defects, their number of defects found may be the same but the number of occurrences would decrease. For example, if each of three PBR perspectives finds one-third of the defects, the number of defects found is 100% but the number of occurrences for all is only 33.3%.

- **Effectiveness**: the average percentage of defects found by subjects from each group, it is calculated as following:

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

in which x_i is the number of defects found by 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 number of defects found by each subject per hour, it is calculated as following:

$$(\sum_{i=1}^{n}(x_i / k)) / n$$

in which x_i is the number of defects found by subject i, k is the total time (in hours) used by each subject to detect the defects and n is the number of subjects in the group;