LUND UNIVERSITY

Are the perspectives really different? Further experimentation on scenario-based reading of requirements

Regnell, Björn; Runeson, Per; Thelin, Thomas

Published in:
Empirical Software Engineering

DOI:
10.1023/A:1009848320066

2000

Link to publication

Citation for published version (APA):
Regnell, B., Runeson, P., & Thelin, T. (2000). Are the perspectives really different? Further experimentation on scenario-based reading of requirements. *Empirical Software Engineering*, *5*(4), 331-356. https://doi.org/10.1023/A:1009848320066

Total number of authors:
3

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Are the Perspectives Really Different?

- Further Experimentation on Scenario-Based Reading of Requirements

Björn Regnell, Per Runeson, Thomas Thelin

Department of Communication Systems, Lund University

Keywords: Requirements Inspection, Perspective-Based Reading, Controlled Experiment

Contact author:

Björn Regnell
Dept. of Communication Systems
Lund University
P.O. Box 118
SE-221 00 LUND
Sweden

Email: bjorn.regnell@telecom.lth.se
Telephone: +46 46 222 90 09
Telefax: +46 46 14 58 23

Are the Perspectives Really Different? – Further Experimentation on Scenario-Based Reading of Requirements

Abstract

Perspective-Based Reading (PBR) is a scenario-based inspection technique where several reviewers read a document from different perspectives (e.g. user, designer, tester). The reading is made according to a special scenario, specific for each perspective. The basic assumption behind PBR is that the perspectives find different defects and a combination of several perspectives detects more defects compared to the same amount of reading with a single perspective. This paper presents a study which analyses the differences in the perspectives. The study is a partial replication of previous studies. It is conducted in an academic environment using graduate students as subjects. Each perspective applies a specific modelling technique: use case modelling for the user perspective, equivalence partitioning for the tester perspective and structured analysis for the design perspective. A total of 30 subjects were divided into 3 groups, giving 10 subjects per perspective. The analysis results show that (1) there is no significant difference among the three perspectives in terms of defect detection rate and number of defects found per hour, (2) there is no significant difference in the defect coverage of the three perspectives, and (3) a simulation study shows that 30 subjects is enough to detect relatively small perspective differences with the chosen statistical test. The results suggest that a combination of multiple perspectives may not give higher coverage of the defects compared to single-perspective reading, but further studies are needed to increase the understanding of perspective difference.

1. Introduction

The validation of requirements documents is often done manually, as requirements documents normally include informal representations of what is required of an intended software system. A commonly used technique for manual validation of software documents is inspections, proposed by Fagan (1976). Inspections can be carried out in different ways and used throughout the software development process for (1) understanding, (2) finding defects, and (3) as a basis for making decisions. Inspections are used to find defects early in the development process, and have shown to be cost effective (e.g. Doolan, 1992).

A central part of the inspection process is the *defect detection* carried out by an individual reviewer reading the document and recording defects (a part of preparation, see Humphrey, 1989). Three common techniques for defect detection are Ad Hoc, Checklist and Scenario-based reading (Porter *et al.*, 1995). Ad Hoc detection denotes an unstructured technique which provides no guidance, implying that reviewers detect defects based on their personal knowledge and experience. The checklist detection technique provides a list of issues and questions, capturing the knowledge of previous inspections, helping the reviewers to focus their reading.

In the scenario-based approach, different reviewers have different responsibilities and are guided in their reading by specific scenarios which aim at constructing a model, instead of just passive reading.

A scenario[1] here denotes a script or procedure that the reviewer should follow. Two variants of scenario-based reading have been proposed: Defect-Based Reading (Porter *et al.*, 1995) and Perspective-Based Reading (Basili *et al.*, 1996). The former (subsequently denoted DBR) concentrates on specific defect classes, while the latter (subsequently denoted PBR) focuses on the points of view of the users of a document.

Another part of the inspection process is the *compilation of defects* into a consolidated defect list where all individual reviewers' defect lists are combined. This step may include the removal of false positives (reported defects that were not considered to be actual defects) as well as the detection of new defects. This step is often done in a structured *inspection meeting* where a *team* of reviewers participate. The effectiveness of the team meeting has been questioned and studied empirically by Votta (1993) and Johnson and Tjahjono (1998).

This paper describes research on scenario-based reading with a PBR approach. The research method is empirical and includes a formal factorial experiment in an academic environment. The presented experiment is a partial replication of previous experiments in the area and focuses on refined hypotheses regarding the differences among the perspectives in PBR. The paper concentrates on defect detection by *individual reviewers*, while the team meeting aspects are not included.

The structure of the paper is as follows. Section 2 gives an overview of related work by summarising results from previously conducted experiments in requirements inspections with a scenario-based approach. Section 3 includes the problem statement motivating the presented work. In Section 4, the experiment plan is described including a discussion on threats to the validity of the study, and Section 5 reports on the operation of the experiment. The results of the analysis is given in Section 6, and Section 7 includes an interpretation of the results. Section 8 provides a summary and conclusions.

2. Related Work

The existing literature on empirical software engineering includes a number of studies related to inspections, where formal experimentation has shown to be a relevant research strategy (Wohlin *et al.*, 2000). The experiment presented in this paper relates to previous experiments on inspections with a scenario-based approach. The findings of a number of experiments on scenario-based inspection of requirements documents are summarized below.

1. The *Maryland-95* study (Porter *et al.*, 1995) compared DBR with Ad Hoc and Checklist in an academic environment. The experiment was run twice with 24 subjects in each run. The requirements documents used were a water level monitoring system (WLMS, 24 pages) and an automobile cruise control

1. There is considerable risk for terminology confusion here, as the term *scenario* also is used within requirements engineering to denote a sequence of events involved in an envisaged usage situation of the system under development. A *use case* is often said to cover a set of related (system usage) scenarios. In scenario-based reading, however, the term scenario is a meta-level concept, denoting a procedure that a reader of a document should follow during inspection.

system (CRUISE, 31 pages).
Result 1: DBR reviewers have significantly higher defect detection rates than either Ad Hoc or Checklist reviewers.
Result 2: DBR reviewers have significantly higher detection rates for those defects that the scenarios were designed to uncover, while all three methods have similar detection rates for other defects.
Result 3: Checklist reviewers do *not* have significantly higher detection rates than Ad Hoc reviewers.
Result 4: Collection meetings produce *no* net improvement in the detection rate – meeting gains are offset by meeting losses.

2. The *NASA* study (Basili *et al.*, 1996) compared PBR with Ad Hoc in an industrial environment. The experiment consisted of a pilot study with 12 subjects and a second main run with 13 subjects. There were two groups of requirements documents used; general requirements documents: an automatic teller machine (ATM, 17 pages), a parking garage control system (PG, 16 pages); and two flight dynamics requirements documents (27 pages each).
Result 1: Individuals applying PBR to general documents have significantly higher detection rates compared to Ad Hoc.
Result 2: Individuals applying PBR to NASA-specific documents do *not* have significantly higher detection rates compared to Ad Hoc.
Result 3: Simulated teams applying PBR to general documents have significantly higher detection rates compared to Ad Hoc.
Result 4: Simulated teams applying PBR to NASA-specific documents have significantly higher detection rates compared to Ad Hoc.
Result 5: Reviewers with more experience do *not* have higher detection rates.

3. The *Kaiserslautern* study (Ciolkowski *et al.*, 1997) compared PBR with Ad Hoc in an academic environment using the ATM and PG documents from the NASA study. The experiment consisted of two runs with 25 and 26 subjects respectively.
Result 1: Individuals applying PBR to general documents have significantly higher detection rates compared to Ad Hoc.
Result 2: Simulated teams applying PBR to general documents have significantly higher detection rates compared to Ad Hoc.
Result 3: The detection rates of five different defect classes are *not* significantly different among the perspectives.

4. The *Bari* study (Fusaro *et al.*, 1997) compared DBR with Ad Hoc and Checklist in an academic environment using the WLMS and CRUISE documents from the Maryland-95 study. The experiment had one run with 30 subjects.
Result 1: DBR did *not* have significantly higher defect detection rates than either Ad Hoc or Checklist.
Result 2: DBR reviewers did *not* have significantly higher detection rates for those defects that the scenarios were designed to uncover, while all three methods had similar detection rates for other defects.

Result 3: Checklist reviewers did *not* have significantly higher detection rates than Ad Hoc reviewers.
Result 4: Collection meetings produced *no* net improvement in the detection rate – meeting gains where offset by meeting losses.

5. The *Trondheim* study (Sørumgård, 1997) compared the NASA study version of PBR with a modified version of PBR (below denoted PBR2) where reviewers were given more instructions on how to apply perspective-based reading. The study was conducted in an academical environment using the ATM and PG documents from the NASA study. The experiment consisted of one run with 48 subjects.
   Result 1: PBR2 reviewers did *not* have significantly higher defect detection rates than PBR.
   Result 2: Individuals applying PBR2 reviewed significantly longer time compared to those who applied PBR.
   Result 3: Individuals applying PBR2 suggested significantly fewer potential defects compared to those who applied PBR.
   Result 4: Individuals applying PBR2 had significantly lower productivity and efficiency than those who applied PBR.

6. The *Strathclyde* study (Miller *et al.*, 1998) compared DBR with Checklist in an academic environment using the WLMS and CRUISE documents from the Maryland study. The experiment consisted of one run with 50 subjects.
   Result 1: In the WLMS document, DBR did *not* have significantly higher defect detection rates than Checklist.
   Result 2: In the CRUISE document, DBR had significantly higher defect detection rates than Checklist.
   Result 3: Collection meetings produced *no* net improvement in the detection rate – meeting gains were offset by meeting losses.

7. The *Linköping* study (Sandahl *et al.*, 1998) compared DBR with Checklist in an academic environment using the WLMS and CRUISE documents from the Maryland study. More defects were added to the list of total defects. The experiment consisted of one run with 24 subjects.
   Result 1: DBR reviewers did *not* have significantly higher defect detection rates than Checklist reviewers.
   Result 2: DBR reviewers did *not* have significantly higher detection rates than Checklist reviewers.

8. The *Maryland-98* study (Shull, 1998) compared PBR with Ad Hoc in an academic environment using the ATM and PG documents from the Maryland study. The experiment consisted of one run with 66 subjects.
   Result 1: PBR reviewers had significantly higher defect detection rates than Ad Hoc reviewers.
   Result 2: Individuals with high experience applying PBR did *not* have significantly[2] higher defect detection rates compared to Ad Hoc.
   Result 3: Individuals with medium experience applying PBR had significantly higher defect detection rates compared to Ad Hoc.

Result 4: Individuals with low experience applying PBR had significantly higher defect detection rates compared to Ad Hoc.
Result 5: Individuals applying PBR had significantly lower productivity compared to those who applied Ad Hoc.

9. The *Lucent* study (Porter and Votta, 1998) replicated the Maryland-95 study in an industrial environment using 18 professional developers at Lucent Technologies. The replication was successful and completely corroborated the results from the Maryland-95 study.

The results of the different studies vary substantially. An attempt to systematically address the combined knowledge, gained from experiments and replications is reported by Hayes (1999), where *meta-analysis* is applied to the results of the Maryland-95, Bari, Strathclyde, Linköping and Lucent studies. It is concluded from the meta-analysis that the effect sizes for the inspection methods are inhomogeneous across the experiments. The Maryland-95 and Lucent studies show most similar results, and an interpretation of the meta-analysis identifies characteristics which make them different from the other three studies:(1) they are conducted in a context where the subjects are more familiar with the notation used, (2) they are conducted in the US where cruise control are more common in cars than in Europe where the other three studies are performed. These hypotheses are, however, not possible to test with the given data, and thus more experimentation is needed.

Table 1 includes a summary of the presented studies. The Maryland-95, NASA, Kaiserslautern, Maryland-98, and Lucent studies indicate that a scenario-based approach gives higher detection rate. The Bari, Strathclyde, and Linköping studies could, however, not corroborate these results, which motivates further studies to increase the understanding of scenario-based reading.

Table 1. *Summary of studies.*

Study	**Purpose**	**Environment**	**Subjects**	**Significant?**
Maryland-95	DBR vs. AdHoc and Checklist	Academic	24+24	YES
Bari	DBR vs. AdHoc and Checklist	Academic	30	NO
Stratchclyde	DBR vs. Checklist	Academic	50	Inconclusive
Linköping	DBR vs. Checklist	Academic	24	NO
Lucent	DBR vs. AdHoc and Checklist	Industrial	18	YES
NASA	PBR vs. AdHoc	Industrial	12+13	YES
Kaiserslautern	PBR vs. AdHoc	Academic	25+26	YES
Trondheim	PBR vs. PBR2	Academic	48	NO
Maryland-98	PBR vs. AdHoc	Academic	66	YES

2. Results 2-4 of the Maryland-98 study apply a significance level of 0.10, while 0.05 is the chosen significance level in all other results.

Many of the studies concluded that real team meetings were ineffective in terms of defect detection. (There may of course be other good reasons for conducting team meetings apart from defect detection, such as consensus building, competence sharing, and decision making.)

The study presented here is subsequently denoted the *Lund* study. The Lund study is a partial replication of the NASA study, and is based on a lab package (Basili *et al.*, 1998) provided by the University of Maryland in order to support empirical investigations of scenario-based reading. The problem statement motivating the Lund study is given in the subsequent section.

3. Research Questions

The previous studies, summarised in Section 2, have mainly concentrated on comparing scenario-based reading with checklist and Ad Hoc techniques in terms of defect detection rates. The objective of the Lund study is, however, to investigate the basic assumption behind scenario-based reading, that the different perspectives find different defects. Another interest is the efficiency of the different perspectives in terms of defects detected per hour. The following two questions are addressed:

1. Do the perspectives detect different defects?
2. Is one perspective superior to another?

There are two aspects of superiority that are addressed: *effectiveness*, i.e. how high fraction of the existing defects are found (detection rate), and *efficiency*, i.e. how many defects are found per time unit.

The perspectives proposed by Basili *et al.* (1996) are designer, tester and user. The users are important stakeholders in the software development process, and especially when the requirements are elicited, analysed and documented. The user role in PBR is focused on detecting defects at a high abstraction level related to system usage, while the designer is focused on internal structures and the tester is focused on verification.

Previous studies have mainly concentrated on the effectiveness in terms of detection rate. From a software engineering viewpoint it is important also to assess the efficiency (e.g. in terms of detected defects per time unit), as this factor is important for a practitioner's decision to introduce a new reading technique. The specific project and application domain constraints then can, together with estimations of how much effort is needed, be a basis for a trade-off between quality and cost.

One main purpose of PBR is that the perspectives detect different kinds of defects in order to minimise the overlap among the reviewers. Hence, a natural question is whether reviewers do find different defects or not. If they detect the same defects, the overlap is not minimised and PBR does not work as it was meant to. If all perspectives find the same kinds of defects it may be a result of (1) that the scenario-based reading approach is inappropriate, (2) that the perspectives may be insufficiently supported by their accompanying scenarios, or (3) that other perspectives are needed to gain a greater coverage difference. The optimal solution is to

use perspectives with no overlap and as high defect detection rate as possible, making PBR highly dependable and effective. The Lund study addresses the overlap by investigating whether the perspectives detect different defects.

Research question 1 is also interesting from a defect content estimation perspective. The capture-recapture approach to defect content estimation uses the overlap among the defects that the reviewers find to estimate the number of remaining defects in a software artifact (Eick *et al.*, 1992; Miller, 1999). The robustness of capture-recapture using PBR is studied by Thelin and Runeson (1999), with the aim of investigating capture-recapture estimators applied to PBR inspections under the hypothesis that PBR works according to its underlying assumption. In the Lund study it is investigated whether the assumptions of PBR are factual. Hence, the Lund study and the Thelin and Runeson (1999) study complement each other in order to answer the question whether capture-recapture estimations can be used for PBR inspections.

4. Experiment Planning

This section describes the planning of the reading experiment. The planning includes the definition of dependent and independent variables, hypotheses to be tested in the experiment, experiment design, instrumentation and an analysis of threats to the validity of the experiment (Wohlin *et al.*, 2000).

The reading experiment is conducted in an academical environment with close relations to industry. The subjects are fourth-year students at the Master's programmes in Computer Science & Engineering and Electrical Engineering at Lund University.

4.1 Variables

The independent variables determine the cases for which the dependent variables are sampled. The purpose is to investigate different reading perspectives and methods, applied to two objects (requirements documents). The inspection objects are the same as in the University of Maryland lab package (Basili *et al.*, 1998), and the design and instrumentation are also based on this lab package. The variables in the study are summarized in Table 2 together with brief explanations.

4.2 Hypotheses

Perspective-Based Reading is assumed to provide more efficient inspections, as different reviewers take different perspectives making the defect overlap smaller (Basili *et al.*, 1996). The objective of the study is to empirically test whether these assumptions are true. In consequence, hypotheses related to performance of different perspectives are stated below. The three null hypotheses address efficiency, effectiveness and distribution over perspectives.

Table 2. *Variables.*

	Name	Values	Description
Independent variables	PERSP	{U,T,D}	One of three perspectives is applied by each subject: User, Tester, and Designer.
	DOC	{ATM,PG}	The inspection objects are two requirements documents: one for an automatic teller machine (ATM) and one for a parking garage control system (PG). The ATM document is 17 pages and contains 29 defects. The PG document is 16 pages and contains 30 defects.
Controlled Variable	EXPERIENCE	Ordinal	The experience with user, tester, design perspectives is measured on a five-level ordinal scale and used in the allocation of subjects to perspectives. (See Sections 4.3 and 6.4)
Dependent Variables	TIME	Integer	The time spent by each reviewer in individual preparation is recorded by all subjects. The time unit used is minutes.
	DEF	Integer	The number of defects found by each reviewer is recorded, excluding false positives. The false positives are removed by the experimenters, in order to ensure that all defect candidates are treated equally.
	EFF	60*DEF/TIME	The defect finding efficiency, i.e. the number of defects found per hour, is calculated as (DEF*60)/TIME.
	RATE	DEF/TOT	The defect finding effectiveness, i.e. the fraction of found defects by total number of defects (also called detection rate) is calculated as DEF divided by the total number of known defects contained in the inspected documents.
	FOUND	Integer	The number of reviewers belonging to a certain perspective, which have found a certain defect in a specific document is recorded. This variable is used for analysing defect finding distributions for different perspectives.

- $H_{0,EFF}$. The perspectives are assumed to have the same finding efficiency, i.e. the number of defects found per hour of inspection is not different for the various perspectives.
- $H_{0,RATE}$. The perspectives are assumed to have the same effectiveness or detection rates, i.e. the fraction of defects identified is not different for the various perspectives.
- $H_{0,FOUND}$. The perspectives are assumed to find the same defects, i.e. the distributions over defects found are the same for the different perspectives.

4.3 Design

To test these hypotheses an experiment with a factorial design (Montgomery, 1997) is used with two factors (PERSP and DOC). The design is summarized in Table 3. The experiment varies the three perspectives over two documents.

Table 3. *Experiment design*

		PERSP		
		User	**Designer**	**Tester**
DOC	ATM	5	5	5
	PG	5	5	5

The assignment of an individual subject to one of the three PBR perspectives (U, D, T), was conducted based on their reported experience (see Section 6.4), similar to the NASA study (Basili *et al.*, 1996). The objective of experience-based perspective assignment is to ensure that each perspective gets a fair distribution of experienced subjects, so that the outcome of the experience is affected by perspective difference rather than experience difference. The experience questionnaire required the subjects to grade their experience with each perspective on a five level ordinal scale. The subjects were then sorted three times, giving a sorted list of subjects for each perspective with the most experienced first. Within the same experience level, the subjects were placed in random order. The subjects were then assigned to perspectives by selecting a subject on top of a perspective list and removing this subject in the other lists before continuing with the next perspective in a round robin fashion starting with a randomly selected perspective, until all subjects were assigned a perspective.

The instruments of the reading experiment consist of two requirements documents and reporting templates for time and defects. These instruments are taken from the University of Maryland lab package (Basili *et al.*, 1998) and are reused with minimal changes.

The factorial design described above is analysed with descriptive statistics (bar plots and box plots) and analysis of variance (ANOVA) (Montgomery, 1997) for the hypotheses $H_{0,EFF}$, and $H_{0,RATE}$.

For the $H_{0,FOUND}$ hypothesis a Chi-square test (Siegel and Castellan, 1988) is used together with a correlation analysis (Robson, 1993).

4.4 Threats to Validity

The validity of the results achieved in experiments depends on factors in the experiment settings. Different types of validity can be prioritized depending on the goal of the experiment. In this case, threats to four types of validity are analysed (Cook and Campbell, 1979; Wohlin *et al.*, 2000): conclusion validity, internal validity, construct validity and external validity.

Conclusion validity concerns the statistical analysis of results and the composition of subjects. In this experiment, well known statistical techniques are applied

which are robust to violations of their assumptions. One general threat to conclusion validity is, however, the low number of samples, which may reduce the ability to reveal patterns in the data. In particular, there are few samples for the Chi-square test, which is further elaborated in Section 6.3.

Internal validity concerns matters that may affect the independent variable with respect to causality, without the researchers knowledge. There are two threats to internal validity in this experiment, selection and instrumentation. The experiment was a mandatory part of a software engineering course, thus the selection of subjects is not random, which involves a threat to the validity of the experiment. The requirements documents used may also affect the results. The documents are rather defect-prone and additional issues in the documents could be considered as defects. On the other hand, it is preferable to have the same definition of defects as in the previous studies for comparison reasons. Other threats to internal validity are considered small. Each subject was only allocated to a single object and a single treatment, hence there is no threat of maturation in the experiment. The subjects applied different perspectives during inspection, but the difference among perspectives are not large enough to suspect compensatory equalisation of treatments or compensatory rivalry. The subjects were also told that their grading in the course was not depending on their performance in the experiment, only on their serious attendance. There is of course a risk that the subjects lack motivation; they may, for example, consider their participation a waste of time or they may not be motivated to learn the techniques. The teacher in the course in which the experiment was performed has, however, made a strong effort in motivating the students. It was clearly stated that a serious participation was mandatory for passing the course. It is the teacher's opinion that the students made a very serious attempt in their inspection.

Construct validity concerns generalisation of the experiment result to concept or theory behind the experiment. A major threat to the construct validity is that the chosen perspectives or the reading techniques for the perspectives may not be representative or good for scenario-based reading. This limits the scope for the conclusions made to these particular perspectives and techniques. Other threats to the construct validity are considered small. The subjects did not know which hypotheses were stated, and were not involved in any discussion on advantages and disadvantages of PBR, thus they were not able to guess what the expected results were.

External validity concerns generalisation of the experiment result to other environments than the one in which the study is conducted. The largest threat to the external validity is the use of students as subjects. However, this threat is reduced by using fourth-year students which are close to finalise their education and start working in industry. The setting is intended to resemble a real inspection situation, but the process that the subjects participate in is not part of a real software development project. The assignments are also intended to be realistic, but the documents are rather short, and real software requirements documents may include many more pages. The threats to external validity regarding the settings and assignments are, however, considered limited, as both the inspection process and the documents resemble real cases to a reasonable extent.

It can be concluded that there are threats to the construct, internal and external validity. However, these are almost the same as in the original studies. Hence, as

long as the conclusions from the experiment are not drawn outside the limitations of these threats, the results are valid.

5. Experiment Operation

The experiment was run during spring 1998. The students were all given a two hour introductory lecture where an overview of the study was given together with a description of the defect classification. A questionnaire on experience was given and each subject was assigned to a perspective, as described in Section 4.3. The students were informed that the experiment was a compulsory part of the course, but the grading was only based on serious participation in the study and not on the individual performance of the students. The anonymity of the students was guaranteed.

A two hour exercise was held, where the three PBR perspectives were described and illustrated using a requirements document for a video rental system (VRS). During the second hour of the exercise, the subjects were practising their own perspective reading technique for the VRS document, and had the opportunity to ask questions. The data collection forms were also explained and used during the exercise. The perspective-based reading of the VRS document was completed by the students on their own after the classroom hours.

The hand-outs for the experiment, which were handed out during the exercise, included the following instrumentation tools:

1. Defect Classification which describes defect classes to be used in the defect list.
2. Time Recording Log for recording the time spent on reading.
3. Defect List for recording the found defects.
4. Reading Instruction, specific for the user, designer, and tester perspectives respectively.
5. Modelling Forms, specific for the user, designer, and tester perspectives respectively.
6. The requirements document (either ATM or PG).

The students were instructed not to discuss the ATM or PG documents and the defects that they find. They were allowed to discuss the PBR perspectives in relation to the VRS document before they started with the actual data collection.

6. Data Analysis

This section presents the statistical analysis of the gathered data. The data were collected from the hand-ins from subjects. Each defect in each subject's defect log was compared with the original "correct" defect list provided by the University of

Maryland lab package. In a meeting, the authors discussed each defect and decided whether it corresponded to a “correct” defect. If no corresponding “correct” defect was found, the reported defect was considered a false positive[3]. The reported time spent was also collected and the EFF, RATE, and FOUND measures were calculated. The total data sets are given in Appendices A and B.

6.1 Individual Performance for Different Perspectives

Box-plots[4] of individual performance in terms of number of defects found per hour (EFF), and the fraction of found defects against the total number of defects (RATE), are shown in Figure 1. The box-plots are split by document and perspective.

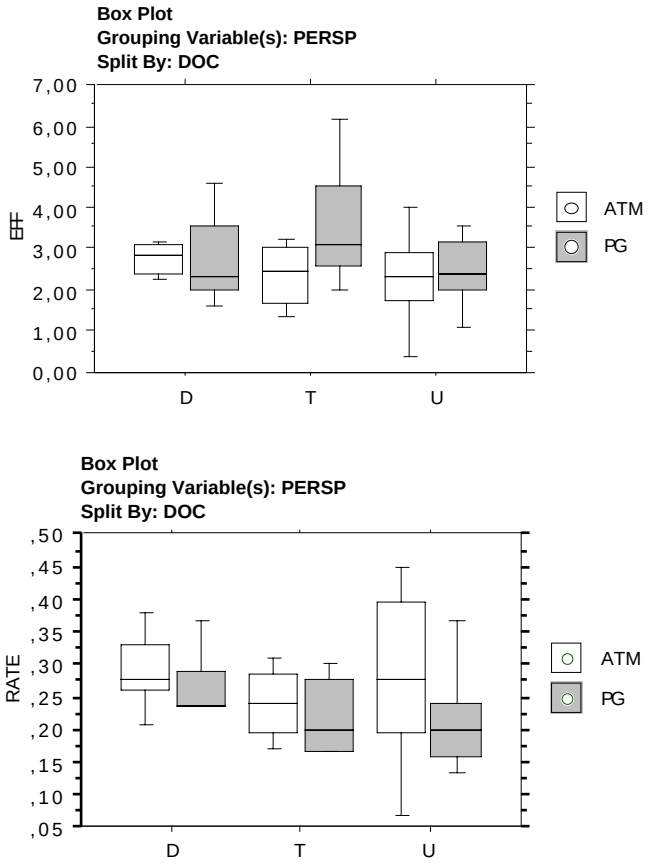

Figure 1. *Box plots for EFF and RATE split by DOC and PERSP.*

3. Some of the defects that where decided to be false positives may in fact be true defects if the defect list from the Maryland lab package is incomplete. It was decided, however, that it is important from a replication viewpoint that the same list of “correct” defects was used. This decision is not considered to have any significant impact on the result as there where only few false positives that were questionable.
4. The box-plots are drawn with the box height corresponding to the 25th and 75th percentile, with the 50th percentile (the median) marked in the box. The whiskers correspond to the 10th and 90th percentile.

ANOVA Table for EFF

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
PERSP	2	1,751	,875	,737	,4893	1,473	,156
DOC	1	1,640	1,640	1,380	,2516	1,380	,193
PERSP * DOC	2	2,229	1,114	,937	,4055	1,875	,187
Residual	24	28,527	1,189				

ANOVA Table for RATE

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
PERSP	2	,012	,006	,802	,4602	1,604	,166
DOC	1	,011	,011	1,488	,2344	1,488	,205
PERSP * DOC	2	,004	,002	,259	,7739	,518	,085
Residual	24	,172	,007				

Figure 2. *ANOVA tables for EFF and RATE.*

For EFF, the Tester perspective on the PG document has a higher mean than the User and Designer perspectives, while for the ATM document, the Designer perspective has a higher mean. For RATE the Designer means are higher compared to the User and Tester perspectives for both documents. There are, however, too few data points per group for any further interpretation of the box-plots, with respect to outliers and skewdness.

When several dependent variables are measured, the multi-variate analysis of variance (MANOVA) can be used to assess if there exists any statistically significant difference in the total set of means. The results of MANOVA tests regarding the effect of PERSP reveal no significance and indicate absence of interaction effects. Furthermore, there are no significant differences in the means of EFF, RATE for the PERSP variable, as shown by the analysis of variance (ANOVA) in Figure 2. From this analysis it can be concluded that the null hypotheses for EFF and RATE can not be rejected for any of the three perspectives.

6.2 Defects Found by Different Perspectives

The hypothesis $H_{0,FOUND}$ regarding the overlap of the found defects among the perspectives, is studied in this section. Descriptive statistics in the form of bar chart plots are shown in Figure 3. For each document the distribution of number of found defects per perspective is shown. There do not seem to be any particular patterns in the different perspective distributions; the defect findings of each perspective seem similarly spread over the defect space. If there had been large differences in the perspective distributions, the bar plot would presumably have groups of defects where one perspective would have a high number of findings while the others would have a low number of findings.

In order to compare the distributions of found defects for each perspective and investigate if there is a significant difference among which defects the perspectives find, a contingency table is created for which a Chi Square test is made (Siegel and Castellan, 1988, pp. 191-194), as shown in Figure 4. The defects that no perspec-

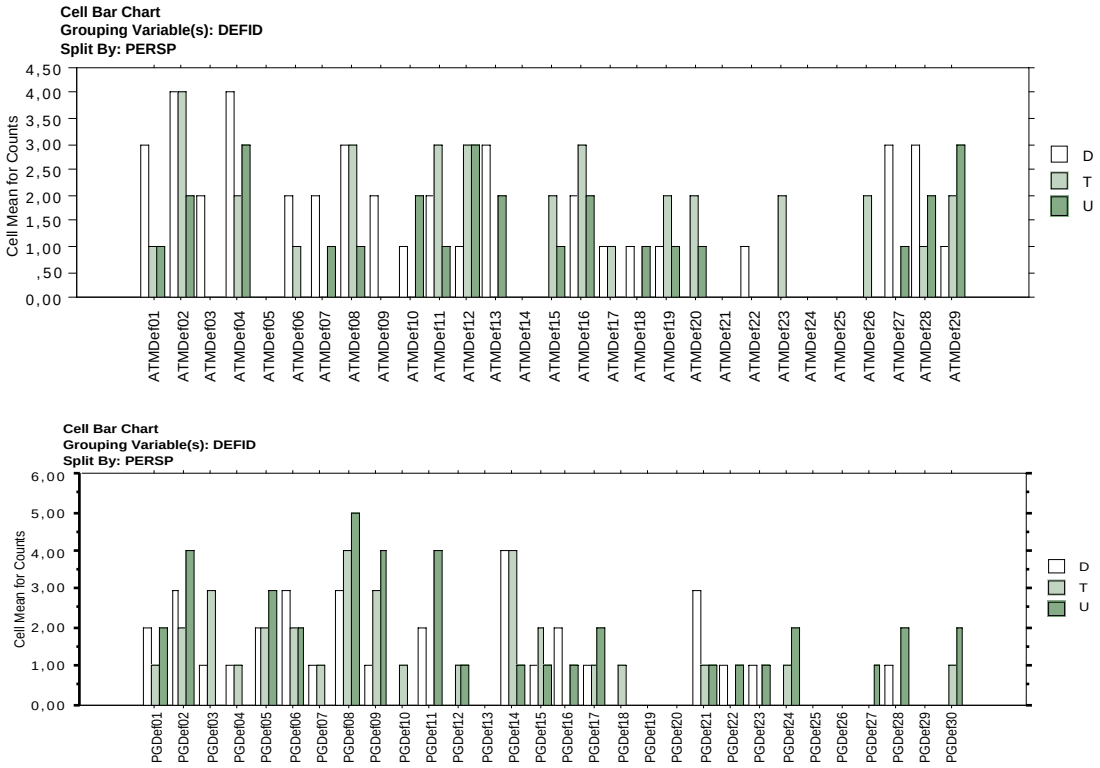

Figure 3. *Bar charts illustrating the distribution of number of reviewers that found each defect.*

tive have found are excluded from the contingency tables (the “Inclusion criteria” in Figure 4), as these cases do not contribute to the testing of differences.

The Chi Square P-values are far from significant, indicating that it is not possible with this test and this particular data set to show a difference in the perspectives’ defect finding distributions.

There are rules of thumb regarding when the Chi Square test can be used (Siegel and Castellan, 1988, pp. 199-200), saying that no more than 20% of the cells should have an expected frequency of less than 5, and no cell should have an expected frequency of less than 1. These rules of thumb are not fulfilled by the data set in this case, but it may be argued that the rules are too conservative and as the expected frequencies in our case are rather evenly distributed, the Chi Square test may still be valid (see further Section 6.3).

The Chi Square test does not give a measure of the *degree* of difference. In order to analyse how different (or similar) the perspectives are, a correlation analysis is presented in Figure 5, using the Pearson correlation coefficient (Robson, 1993, pp. 338-340).

Two different correlation analyses are provided for each document, one with all “correct” defects included and one where only those defects are included that were found by at least one reviewer. The latter may be advocated, as we are interested in the differences in the set of defects that are found by each perspective; the defects that no perspective find do not contribute to differences among perspectives.

Summary Table for DEFID, PERSP
Inclusion criteria: Counts>0 from PG.data

Num. Missing	0
DF	46
Chi Square	33,951
Chi Square P-Value	,9058
G-Squared	•
G-Squared P-Value	•
Contingency Coef.	,494
Cramer's V	,402

Summary Table for DEFID, PERSP
Inclusion criteria: Counts > 0 from ATM.data

Num. Missing	0
DF	46
Chi Square	41,676
Chi Square P-Value	,6538
G-Squared	•
G-Squared P-Value	•
Contingency Coef.	,535
Cramer's V	,448

Observed Frequencies for DEFID, PERSP
Inclusion criteria: Counts>0 from PG.data

	D	T	U	Totals
PGDef01	2	1	2	5
PGDef02	3	2	4	9
PGDef03	1	3	0	4
PGDef04	1	1	0	2
PGDef05	2	2	3	7
PGDef06	3	2	2	7
PGDef07	1	1	0	2
PGDef08	3	4	5	12
PGDef09	1	3	4	8
PGDef10	0	1	0	1
PGDef11	2	0	4	6
PGDef12	0	1	1	2
PGDef14	4	4	1	9
PGDef15	1	2	1	4
PGDef16	2	0	1	3
PGDef17	1	1	2	4
PGDef18	0	1	0	1
PGDef21	3	1	1	5
PGDef22	1	0	1	2
PGDef23	1	0	1	2
PGDef24	0	1	2	3
PGDef27	0	0	1	1
PGDef28	1	0	2	3
PGDef30	0	1	2	3
Totals	33	32	40	105

Observed Frequencies for DEFID, PERSP
Inclusion criteria: Counts > 0 from ATM.data

	D	T	U	Totals
ATMDef01	3	1	1	5
ATMDef02	4	4	2	10
ATMDef03	2	0	0	2
ATMDef04	4	2	3	9
ATMDef06	2	1	0	3
ATMDef07	2	0	1	3
ATMDef08	3	3	1	7
ATMDef09	2	0	0	2
ATMDef10	1	0	2	3
ATMDef11	2	3	1	6
ATMDef12	1	3	3	7
ATMDef13	3	0	2	5
ATMDef15	0	2	1	3
ATMDef16	2	3	2	7
ATMDef17	1	1	0	2
ATMDef18	1	0	1	2
ATMDef19	1	2	1	4
ATMDef20	0	2	1	3
ATMDef22	1	0	0	1
ATMDef23	0	2	0	2
ATMDef26	0	2	0	2
ATMDef27	3	0	1	4
ATMDef28	3	1	2	6
ATMDef29	1	2	3	6
Totals	42	34	28	104

Figure 4. *Chi Square tests and contingency tables for defects found by U,T,D per DOC.*

The P-value indicates if the correlation coefficient is significant, and the confidence intervals presented indicate the range wherein the correlation coefficient is likely to be.

The correlation analysis indicates that there are significantly positive correlations among the perspectives, meaning that when one perspective finds a defect it is likely that others also find it. The only correlation coefficient that is far from significant is the Designer-Tester correlation for the ATM document.

Another way of qualitatively analysing the overlap among the perspectives is Venn-diagrams, as used in the NASA study (Basili *et al.*, 1996, p.151).

For the purpose of comparison we include such diagrams for the Lund study data, as shown in Figure 6. Each defect is categorised in one of seven classes depending on which combinations of perspectives that have a FOUND measure greater than zero. The numbers in the Venn-diagrams indicate how many defects

ATM Document

Correlation Analysis

	Correlation	P-Value	95% Lower	95% Upper
User, Tester	,480	,0076	,138	,720
User, Designer	,499	,0052	,162	,732
Tester, Designer	,258	,1789	-,120	,570

29 observations were used in this computation.

Correlation Analysis
Inclusion criteria: User > 0 OR Tester > 0 OR Designer > 0 from ATM-ctable.data

	Correlation	P-Value	95% Lower	95% Upper
User, Tester	,357	,0867	-,054	,665
User, Designer	,352	,0915	-,059	,662
Tester, Designer	,043	,8449	-,367	,439

24 observations were used in this computation.

PG Document

Correlation Analysis

	Correlation	P-Value	95% Lower	95% Upper
User, Tester	,463	,0092	,123	,706
User, Designer	,543	,0016	,228	,756
Tester, Designer	,601	,0003	,307	,790

30 observations were used in this computation.

Correlation Analysis
Inclusion criteria: User > 0 OR Tester > 0 OR Designer > 0 from PG-ctable.data

	Correlation	P-Value	95% Lower	95% Upper
User, Tester	,319	,1300	-,097	,640
User, Designer	,414	,0438	,012	,700
Tester, Designer	,493	,0134	,112	,748

24 observations were used in this computation.

Figure 5. *Correlation analysis of the perspectives for each document.*

that belong to each class. For example, for the PG document, there are 10 defects which were found by all perspectives, while 5 defects were found by both the user and designer perspectives and only one defect was found solely by the user perspective.

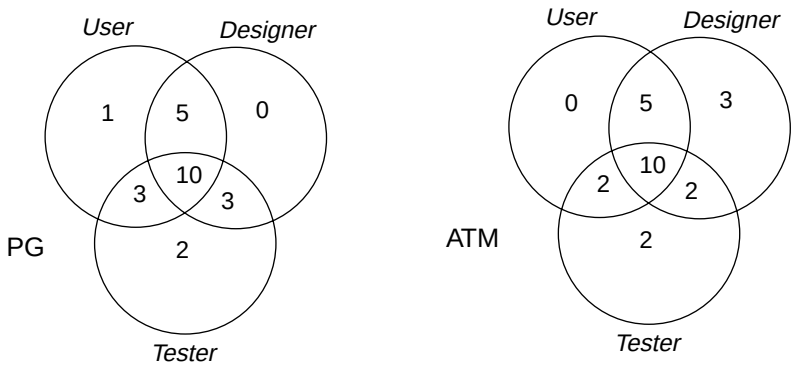

Figure 6. *Defect coverage for the PG and ATM documents.*

This type of analysis is very sensitive to the number of subjects. It is enough that only one reviewer finds a defect, for the classification to change. The probabil-

ity that a defect is found increases with the number of reviewers, and if we have a large number of reviewers, the defects will be more likely to be included in the class where all perspectives have found it. This means that this type of analysis is not very robust, and does not provide meaningful interpretations in the general case. In our case, we can at least say that the defect coverage analysis in Figure 6 does not contradict our previous results that we cannot reject the hypothesis that the perspectives are similar with respect the sets of defects that they find. The defects found by all perspectives is by far the largest class.

6.3 Is the Sample Size Large Enough?

The outcome of the Lund study is that no significant difference among the perspectives can be detected. A question arises whether this is due to lack of differences in the data, or that the statistical tests are not able to reveal the differences, for example, due to the limited amount of data. In order to evaluate the Chi-square test the perspective defect detection data sets are simulated with stochastic variations among perspectives and the Chi-square test is applied to the simulated data.

The simulation is designed to resemble the experiment presented in the previous section. The difference is that in the simulation case, the probability for detection of a specific defect by a perspective is an independent variable. Furthermore, only the FOUND dependent variable is applied, since the time aspect is not modelled. The simulation model is designed as follows:

- The number of defects in each simulated document is 30.
- For every simulated inspection, three perspectives are used with 10 reviewers per perspective. It is assumed that a document contains three different types of defects, which have different probabilities of being detected. One perspective has high probability (P_{HIGH}) to detect one third of the defects and low probability (P_{LOW}) to detect the other two thirds of the defects. The difference between P_{HIGH} and P_{LOW} is denoted P_{Δ}. The probability levels are set to values between 0.05 and 0.5 in steps of 0.05, which are values around the measured mean in the Lund study.
- 1000 runs of each inspection are simulated.

The $H_{0,FOUND}$ hypothesis is tested with the Chi-Square test and the results are presented in Figure 7. Each simulated experiment is tested separately. The figure shows the fraction of tests that are rejected for each case. For all simulation cases with P_{Δ} larger than 0.3, the test can significantly show a difference among the simulated perspectives. For simulation cases with P_{HIGH} lower than 0.25, the differences can be shown if P_{Δ} is larger than 0.2. The tests are conducted with a significance level of 0.05. The simulation study shows that differences in FOUND are possible to detect with the Chi-Square test, even if the perspective differences are small and the sample size is small.

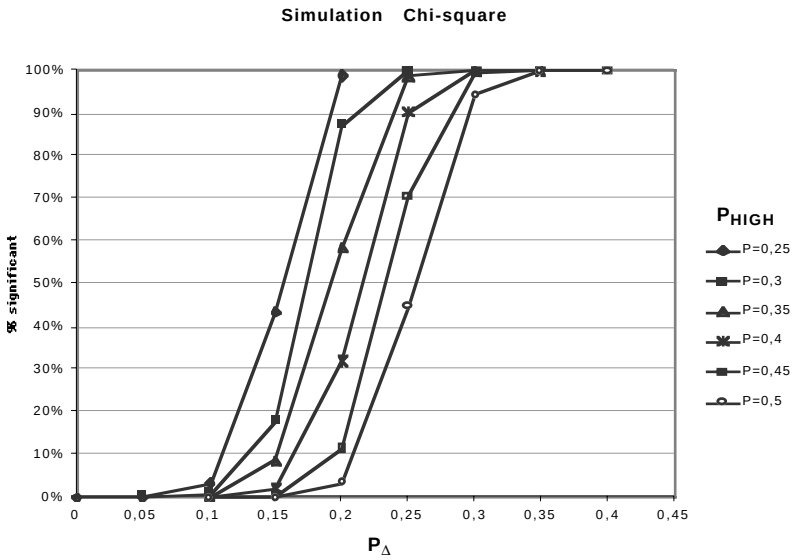

Figure 7. *Fraction of significant test results concerning $H_{0,FOUND}$.*

6.4 Experience of Subjects

The experience was measured through a questionnaire which covers each perspective in general, as well as experience with the specific modelling techniques of the three perspectives (use case modelling, equivalence partitioning, and structured analysis). The experience is measured for each general perspective and each specific modelling technique on a five level ordinal scale: 1 = none, 2 = studied in class or from book, 3 = practised in a class project, 4 = used on one project in industry, 5 = used on multiple projects in industry.

Figure 8 shows the average experience for each subject regarding the perspective to which the subject was assigned, both for the perspective in general and for the specific modelling technique.

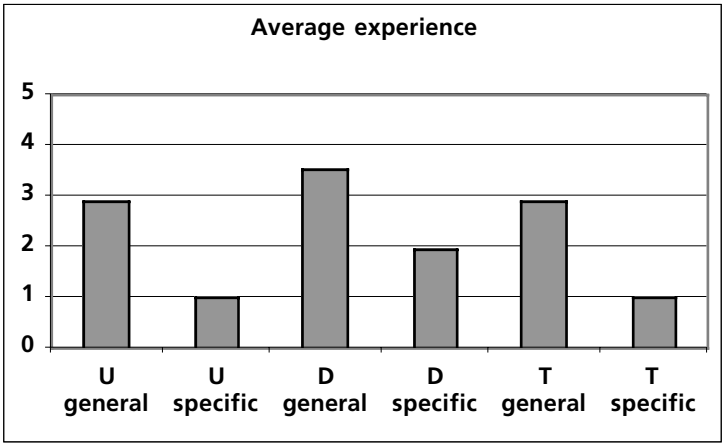

Figure 8. *Average experience of subjects regarding their general experience of their perspective and specific experience with their modelling technique.*

It can be seen that the allocation of subjects (according to the algorithm explained in Section 4.3) has, as expected, resulted in a relatively balanced experience profile over the perspectives. It can also be noted that the students had very little industrial experience.

7. Interpretations of Results

In this section the data analysis is interpreted with respect to the hypotheses stated in Section 4.2. The first two hypotheses are tested using ANOVA and the third hypothesis is tested using a Chi-square test. The following three null-hypotheses can *not* be rejected:

- $H_{0,EFF}$ The perspectives are assumed to find the same number of defects per hour. This hypothesis can *not* be rejected.
- $H_{0,RATE}$ The perspectives are assumed to find the same number of defects. This hypothesis can *not* be rejected.
- $H_{0,FOUND}$. The perspectives are assumed to find the same defects. This hypothesis can *not* be rejected.

It can hence be concluded that there is no significant difference among the three perspectives, user, design and test. This is true for all the three hypotheses, i.e. there is no significant difference in terms of effectiveness or efficiency. Furthermore, there is no significant difference in time spent using the different perspectives, hence, the time spent does not bias in favour of any of the techniques. The lack of difference among the three perspectives does, if the result is possible to replicate and generalize, seriously affect the cornerstones of the PBR. The advantages of PBR are assumed to be that the different perspectives focus on different types of defects, and thus detect different defect sets. This study shows no statistically significant difference among the sets of defects found by the three perspectives, and thus the advantages of PBR can be questioned.

Threats to the conclusion validity of the results are that the number of samples is low, in particular for the Chi-square test. However, a simulation study reveals that the Chi-square test can with 30 subjects detect differences among perspectives for relatively small differences in detection probability. Furthermore, the bar charts over the defects found by different perspectives (see Figure 3) do not indicate any clear pattern, which supports the non-significant results. The ANOVA statistics are applied within acceptable limits, and these do not show any difference among the perspectives. The specific perspectives and the reading techniques for the perspectives might also be a threat to the validity of the results, when trying to apply the results to scenario-based reading in general.

The validity threat regarding the motivation of subjects can be evaluated by comparing the detection rates of the Lund study with other studies. The individual PBR detection rate for the NASA study (Basili *et al.*, 1996) was on average 0,249 for the pilot study and 0,321 for the main run, while the Lund study shows an average individual PBR detection rate of 0,252. The rates are comparable, supporting

the assumption that the subjects in this study was as motivated as in the NASA study.

Other threats to the validity in Section 4.4 are not considered differently in the light of the result.

8. Summary and Conclusions

The study reported in this paper is focused on the evaluation of Perspective Based Reading (PBR) of requirements documents. The study is a partial replication of previous experiments in an academic environment based on the lab package from University of Maryland (Basili *et al.*, 1998).

The objective of the presented study is twofold:

1. Investigate the differences in the performance of the perspectives in terms of effectiveness (defect detection rate) and efficiency (number of found defects per hour).

2. Investigate the differences in defect coverage of the different perspectives, and hence evaluate the basic assumptions behind PBR supposing that different perspectives find different defects.

The experiment setting includes two requirements documents and scenarios for three perspectives (*user* applying use case modelling, *designer* applying structured analysis, and *tester* applying equivalence partitioning). A total of 30 MSc students were divided into 3 groups, giving 10 subjects per perspective.

In summary the results from the data analysis show that:

1. There is no significant difference among the user, designer and tester perspectives in terms of defect detection rate and number of defects found per hour.

2. There is no significant difference in the defect coverage of the three perspectives.

The interpretation of these results suggests that a combination of multiple perspectives may not give higher defect coverage compared to reading with only one perspective.

The results contradict the main assumptions behind PBR. Some of the previous studies, summarized in Section 2, have shown significant advantages with Scenario-based Reading over Ad Hoc inspection, but no statistical analysis on the difference among perspective performance is made in any of the studies reported in Section 2. Furthermore, the previous studies in Section 2 have not taken the efficiency into account (number of defects found per hour), but concentrates on detection rate as the main dependent variable. From a software engineering perspective, where the cost and efficiency of a method are of central interest, it is very interesting to study not only the detection rate, but also if a method can perform well within limited effort.

There are a number of threats to the validity of the results, including:

1. The setting may not be realistic.
2. The perspectives may not be optimal.
3. The subjects may not be motivated or trained enough.
4. The number of subjects may be too small.

It can be argued that the threats to validity are under control, based on the following considerations: (1) The inspection objects are similar to industrial requirements documents; (2) The perspectives are motivated from a software engineering process view; (3) The subjects were 4th year students with a special interest in software engineering attending an optional course which they have chosen out of their own interest, and further, many companies have a large fraction of employees with fresh exams; (4) The presented simulation study shows that relatively small differences among the perspectives can be detected with the chosen analysis for the given number of data points.

A single study, like this, is no sufficient basis for changing the attitudes towards PBR. Conducting the same analyses on data from existing experiments as well as new replications with the purpose of evaluating differences among perspectives will bring more clarity into the advantages and disadvantages of PBR techniques, and also give a better control over the validity threats.

Appendix A. Individual performance

Table 4. *Data for each subject.*

ID	PERSP	DOC	TIME	DEF	EFF	TOT	RATE
1	U	ATM	187	8	2,567	29	0,276
2	D	PG	150	8	3,200	30	0,267
3	T	ATM	165	9	3,273	29	0,310
4	U	PG	185	11	3,568	30	0,367
5	D	ATM	155	8	3,097	29	0,276
6	T	PG	121	8	3,967	30	0,267
7	U	ATM	190	7	2,211	29	0,241
8	D	PG	260	7	1,615	30	0,233
9	T	ATM	123	6	2,927	29	0,207
10	U	PG	155	6	2,323	30	0,200
11	D	ATM	210	11	3,143	29	0,379
12	T	PG	88	9	6,136	30	0,300
13	U	ATM	280	11	2,357	29	0,379
14	D	PG	145	11	4,552	30	0,367
15	T	ATM	170	5	1,765	29	0,172
16	U	PG	120	6	3,000	30	0,200
17	D	ATM	190	9	2,842	29	0,310
18	T	PG	97	5	3,093	30	0,167
19	U	ATM	295	2	0,407	29	0,069
20	D	PG	180	7	2,333	30	0,233
21	T	ATM	306	7	1,373	29	0,241
22	U	PG	223	4	1,076	30	0,133
23	D	ATM	157	6	2,293	29	0,207
24	T	PG	130	6	2,769	30	0,200
25	U	ATM	195	13	4,000	29	0,448
26	D	PG	200	7	2,100	30	0,233
27	T	ATM	195	8	2,462	29	0,276
28	U	PG	125	5	2,400	30	0,167
29	D	ATM	200	8	2,400	29	0,276
30	T	PG	150	5	2,000	30	0,167

Appendix B. Defects found by perspectives

B.1 PG document

Table 5. *Defects id D# found (1) or not found (0) by individuals reading the PG document.*

	Individuals																	
	User Perspective						**Tester Perspective**						**Designer Perspective**					
D#	2	8	14	20	26	S	4	10	16	22	28	S	6	12	18	24	30	S
1	0	0	1	0	1	2	1	0	0	0	0	1	1	0	0	1	0	2
2	1	1	1	0	1	4	1	0	1	0	0	2	0	1	1	1	0	3
3	0	0	0	0	0	0	1	1	0	1	0	3	0	0	0	1	0	1
4	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	1
5	0	0	1	1	1	3	1	0	0	0	1	2	0	1	1	0	0	2
6	1	1	0	0	0	2	0	1	1	0	0	2	1	0	0	1	1	3
7	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	1
8	1	1	1	1	1	5	1	1	1	1	0	4	1	1	0	1	0	3
9	1	1	1	1	0	4	1	1	1	0	0	3	0	1	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
11	1	1	0	1	1	4	0	0	0	0	0	0	0	1	1	0	0	2
12	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	1	1	1	1	0	1	1	4	1	1	0	1	1	4
15	0	0	1	0	0	1	0	1	0	0	1	2	1	0	0	0	0	1
16	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	2
17	0	0	0	1	1	2	1	0	0	0	0	1	0	1	0	0	0	1
18	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	1	0	0	1	0	0	0	0	1	1	1	0	1	0	1	3
22	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1
23	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1
24	0	0	1	1	0	2	0	0	0	0	1	1	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
28	0	1	1	0	0	2	0	0	0	0	0	0	1	0	0	0	0	1
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	1	0	1	0	0	2	0	0	1	0	0	1	0	0	0	0	0	0
S	8	7	11	7	7	40	11	6	6	4	5	32	8	9	5	6	5	33

B.2 ATM document

Table 6. *Defects number D# found (1) or not found (0) by individuals reading the ATM document.*

	Individuals																	
	User Perspective						**Tester Perspective**						**Designer Perspective**					
D#	1	7	13	19	25	S	3	9	15	21	27	S	5	11	17	23	29	S
1	0	0	0	1	0	1	0	0	0	1	0	1	1	1	0	0	1	3
2	1	0	1	0	0	2	1	0	1	1	1	4	1	1	1	0	1	4
3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	2
4	0	1	1	1	0	3	1	1	0	0	0	2	1	1	1	0	1	4
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	2
7	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	2
8	0	0	1	0	0	1	0	1	0	1	1	3	1	1	0	0	1	3
9	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	2
10	0	1	1	0	0	2	0	0	0	0	0	0	0	1	0	0	0	1
11	1	0	0	0	0	1	0	1	0	1	1	3	1	0	1	0	0	2
12	1	1	1	0	0	3	0	0	1	1	1	3	0	1	0	0	0	1
13	1	0	1	0	0	2	0	0	0	0	0	0	0	1	1	1	0	3
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	1	0	0	0	1	1	0	0	0	1	2	0	0	0	0	0	0
16	0	1	1	0	0	2	1	1	1	0	0	3	0	1	1	0	0	2
17	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	1
18	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1
19	1	0	0	0	0	1	1	1	0	0	0	2	0	0	0	1	0	1
20	0	0	1	0	0	1	0	0	1	1	0	2	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
23	0	0	0	0	0	0	0	0	1	0	1	2	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	1	0	1	0	2	0	0	0	0	0	0
27	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	1	3
28	1	0	1	0	0	2	1	0	0	0	0	1	1	0	1	0	1	3
29	1	1	1	0	0	3	1	0	0	0	1	2	0	0	0	1	0	1
S	8	7	11	2	0	28	8	6	5	7	8	34	8	11	9	6	8	42

Acknowledgements

First of all, the authors would like to thank the students who participated as subjects in the experiment. We would also like to give a special acknowledgement to Forrest Shull at University of Maryland who provided support on the UMD lab-pack and gave many good comments on a draft version of this paper. We are also grateful for all constructive comments made by the anonymous reviewers. Thanks also to Claes Wohlin, Martin Höst and Håkan Petersson at Dept. of Communication Systems, Lund University, who have carefully reviewed this paper. Special thanks to Anders Holtsberg at Centre for Mathematical Sciences, Lund University, for his expert help on statistical analysis. This work is partly funded by the National Board of Industrial and Technical Development (NUTEK), Sweden, grant 1K1P-97-09690.

References

Basili, V. R., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sørumgård, S., and Zelkowitz, M. 1996. "The Empirical Investigation of Perspective-Based Reading", *Empirical Software Engineering*, 1(2): 133-164.

Basili, V. R., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sørumgård, S., and Zelkowitz, M. 1998. *Lab Package for the Empirical Investigation of Perspective-Based Reading*
Available at: http://www.cs.umd.edu/projects/SoftEng/ESEG/manual/pbr_package/manual.html

Ciolkowski, M., Differding, C., Laitenberger, O., and Münch, J. 1997. "Empirical Investigation of Perspective-Based Reading: A Replicated Experiment", *ISERN Report* no. 97-13.
Available at: http://www.iese.fhg.de/ISERN/pub/isern_biblio_tech.html

Cook T. D., and Campbell D. T. 1979. *Quasi-Experimentation – Design and Analysis Issues for Field Settings*, Houghton Mifflin Company.

Doolan, E. P. 1992. "Experiences with Fagan's inspection method", *Software Practice and Experience*, 22(2):173-182.

Eick, S. G., Loader, C. R., Long, M. D., Votta, L. G. and Vander Wiel, S. A. 1992. "Estimating Software Fault Content Before Coding", *Proceedings of the 14th International Conference on Software Engineering (ICSE'92)*, pp. 59-65.

Fagan, M. E. 1976. "Design and Code Inspections to Reduce Errors in Program Development", *IBM System Journal*, 15(3):182-211.

Fusaro, P., Lanubile, F., and Visaggio, G. 1997. "A Replicated Experiment to Assess Requirements Inspection Techniques", *Empirical Software Engineering*, 2(1): 39-57.

Hayes, W. 1999. "Research Synthesis in Software Engineering: A Case for Meta-Analysis", *Proceedings of the 6th International Software Metrics Symposium (METRICS'99)*, Boca Raton, Florida, USA, pp. 143-151.

Humphrey, W. S. 1989. *Managing the Software Process*, Addison-Wesley.

Johnson, P. M., and Tjahjono, D. 1998. "Does Every Inspection Really Need a Meeting?", *Empirical Software Engineering*, 3(1): 9-35.

Miller, J. 1999. "Estimating the Number of Remaining Defects after Inspection", *Software Testing, Verification and Reliability*, 9(4):167-189.

Miller, J., Wood, M., and Roper, M. 1998. "Further Experiences with Scenarios and Checklists", *Empirical Software Engineering*, 3(1): 37-64.

Montgomery, D. C. 1997. *Design and Analysis of Experiments*, Fourth Edition. Wiley.

Porter, A., Votta, L., and Basili, V. R. 1995. “Comparing Detection Methods for Software Requirements Inspection: A Replicated Experiment”, *IEEE Transactions on Software Engineering*, 21(6):563-575.

Porter, A., and Votta, L. 1998. “Comparing Detection Methods for Software Requirements Inspection: A Replication Using Professional Subjects”, *Empirical Software Engineering*, 3(4):355-380.

Robson, C. 1993. *Real World Research*. Blackwell.

Sandahl, K., Blomkvist, O., Karlsson, J., Krysander, C., Lindvall, M., and Ohlsson, N. 1998. “An Extended Replication of an Experiment for Assessing Methods for Software Requirements”, *Empirical Software Engineering*, 3(4):381-406.

Shull, F. 1998. *Developing Techniques for Using Software Documents: A Series of Empirical Studies*, PhD Thesis, Computer Science Department, University of Maryland, USA.

Siegel, S., and Castellan N. J. 1988. *Nonparametric Statistics for the Behavioral Sciences*, Second Edition. McGraw-Hill.

Sørumgård, S. 1997. *Verification of Process Conformance in Empirical Studies of Software Development*, PhD Thesis, Department of Computer and Information Science, The Norwegian University of Science and Technology, Norway.

Thelin, T., and Runeson, P. 1999. “Capture-Recapture Estimations for Perspective-Based Reading – A Simulated Experiment”, *Proceedings of the International Conference on Product Focused Software Process Improvement (PROFES'99)*, Oulu, Finland, pp. 182-200.

Votta, L. G. 1993. “Does Every Inspection Need a Meeting?”, *Proceedings of the ACM SIGSOFT 1993 Symposium on Foundations of Software Engineering, ACM Software Engineering Notes*, 18(5):107-114.

Weidenhaupt, K., Pohl, K., Jarke, M., and Haumer, P. 1998. “Scenarios in System Development: Current Practice”, *IEEE Software*, March/April 1998, pp. 34-45.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., Wesslén, A. 2000. *Experimentation in Software Engineering – An Introduction*, Kluwer.