Markus Ciolkowski, Christiane Differding, Oliver Laitenberger, Jürgen Münch, "Empirical Investigation on Perspective-Based Reading: A Replicated Experiment", Technical Report No. 13/97, International Software Engineering Research Network (ISERN), 1997.

Empirical Investigation of Perspective-based Reading: A Replicated Experiment

Marcus Ciolkowski, Christiane Differding, Oliver Laitenberger, Jürgen Münch*
ISERN-Report-97-13

Abstract

Inspection is considered a powerful method to detect defects in a software artifact. It is reported that savings are particularly high if inspections are used in early phases of the software development process, i.e. in the requirements definition phase. However, only few systematic techniques exist to support defect detection in requirements documents. One is perspective-based reading (PBR). The effectiveness of this technique has been validated in an experiment with software professionals at NASA/Goddard Space Flight Center.

In this paper we describe a replication of this experiment within an academic environment to validate the original results. In the original experiment, no real team meetings were performed, i.e. the individual results were pooled into nominal team results. In contrast, in our replication effort we performed real team meetings, which allowed us to make a comparison between real and nominal teams. Moreover, we investigated how the technique supports detection of defects belonging to different defect classes. The results are three-fold: (1) We basically confirm the results of the original experiment, that PBR helps to increase individual and team defect detection effectiveness compared to an Ad-hoc approach. (2) We found no statistically significant difference between real and nominal teams. (3) The analysis according to different defect classes only yield few statistically significant results due to the experimental setting.

To document data collection and analysis, we used the Goal/Question/Metric approach. We found it highly beneficial for structuring data collection and analysis especially for replication purposes.

Keywords: experimental software engineering, inspection, perspective-based reading, reading technique, replication, Goal/Question/Metric approach.

1. Introduction

Defect detection in software artifacts is crucial to the success of software development projects. Defects in a software artifact increase cycle time and development costs if they are not detected and removed before the next development phase begins *[16]*. Hence, it makes sense to start defect detection with requirements documents since it is the basic document the software system is built upon. However, only few methods are available for defect detection in requirements documents.

One method is inspection *[1, 18]*. An inspection process usually consists of three phases: defect detection, defect collection, and defect repair. In most inspection processes, and their associated reading technique, the defect detection phase is not supported. There, individual defect detection depends upon chance or experience factors. Thus, we are focusing our efforts on reading techniques that support individual defect detection. So far, only few reading techniques have been proposed in the literature. Examples are checklists *[19,21]*, Mill's reading by stepwise abstraction for code documents *[27]*, Defect-based Reading *[32]*, or the one we consider in this paper, Perspective-based Reading (PBR) *[7]*.

Little is known about the practical benefits of existing reading techniques, and how effective these techniques are in a given context. One way to gather knowledge about the practical benefits of reading techniques is to evaluate their performance in controlled experiments *[7,11,24,32]*. The number of subjects participating in such experiments is often low, especially if experiments are conducted in an industrial setting. Thus, it is necessary to perform replications of the original experiment in similar or slightly differing experimental settings to increase the confidence in the original findings. Furthermore, replication helps generalize the results, especially when they are conducted in different contexts. We distinguish two forms of replication: internal and external replication *[13]*. Internal replication is undertaken by the original experimenters;

external by independent researchers. Brooks et al. *[13]* state that external replication is critical for establishing sound results and that it provides either supporting evidence or questions the validity of the original experiment.

1.1 Original Experiment

We replicated and extended an experiment that evaluates the effectiveness of the PBR reading technique. The experiment was performed by Basili *et al. [7]*. Throughout the paper we refer to it as the *original experiment*. It consisted of two runs which were performed with professional software developers from the NASA / Goddard Space Flight Center (NASA/GSFC). In each run, the subjects read two kinds of requirements documents using two reading techniques for each kind of document. The reading techniques were PBR and their usual NASA approach that had evolved over several years. The two requirements documents were one generic in nature and the other belonging to the NASA problem domain. The main result of this experiment was that individuals using PBR performed better than using the NASA approach especially when they were less familiar with the domain. Instead of performing real team meetings, individual results were pooled together to "simulate" team meetings. We call the simulated teams "nominal teams" throughout the paper. The nominal teams applying PBR performed in most cases significantly better than the nominal teams applying the NASA approach.

1.2 Replication

This paper describes an external replication of the original experiment at the University of Kaiserslautern, Germany. In the following, we describe the replication as the *experiment*. We conducted two runs of the experiment with students of the Computer Science Department. The subjects read two requirements documents of the generic type using two reading techniques, Perspective-based reading and Ad-hoc reading.

We tried to conserve as much as possible the context of the original experiment. The main differences regarding the context of the experiment concern the subjects (professional developers vs. students) and the inspected documents (only the generic type). The replication was intended to validate the results of the original experiment and to examine further aspects of PBR not yet considered: the benefits of real team meetings and the detection of defects belonging to different defect classes. To improve the support for further replications we used techniques of the Goal/Question/Metric (GQM) approach to document the derivation of measures and the analysis.

1.3 Paper Outline

The paper is organized as follows: Section 2 describes the experiment. Section 3 presents the data analysis results. Section 4 discusses the threats to validity. Finally, Section 5 summarizes the paper and presents areas for future work.

2. Description of the Experiment

In this chapter we explain the theory behind PBR, the goal of the experiment with the hypotheses and the dependent and independent variables. We describe the experimental design and the plan for the analysis procedure. We give an overview over the experimental material, the subjects, and the experimental procedure.

2.1 Perspective-based Reading

Empirical work must result in a deep understanding of the phenomena under study *[28]*. However, before conducting an experiment the researcher relies on a theory that will predict or explain most of the results. Thus, we will explain the theory behind PBR in the next subsection. Although the description of PBR in the original experiment included the most important aspects, we believe that some details were missing that may be important to understand the rationales behind PBR. Thus, we refined the description based on *[7]* and *[25]*.

Theory behind Perspective-based Reading

PBR is a technique to support defect detection in a software product. The idea is to use perspectives that different customers or consumers of the software product have to verify the correctness of the software product. The following goals have been established while developing PBR:

1. PBR should be adaptable to the particular document (e.g. requirements document) and the notation in which the document is written. That is, it should fit the appropriate development phase and notation. The rationale is that PBR should be applicable to every software artifact within the software development lifecycle.
2. PBR should be tailorable, based upon the project and environment characteristics. The rationale is that any technique must be adapted to the particularities of a given environment to be most successful.
3. PBR should be detailed in that it provides the reader with a well-defined reading process. The rationale is that readers get concrete guidelines on how to read a document. Thus, defect detection becomes less dependent on factors like experience. Moreover, the reading process becomes repeatable because PBR provides process guidance for readers.

4. PBR should be focused in that a particular perspective provides particular coverage of part of the document, and a combination of perspectives provides coverage of the entire document. The rationale is that an read does not have to check all details of the inspected document; s/he only checks the details that are important with respect to his/her perspective. However, as a document is usually read by various inspectors, all important characteristics are verified.
5. PBR should be specific in that each inspector has a particular purpose or goal for reading the document and procedures that support that goal. The rationale is that an inspector knows exactly from what perspective to read the document and what to look for.

After defining the goals and rationales we describe the model underlying Perspective-based reading. One major purpose of reading a document is comprehension. Comprehension is a necessary requirement for detecting defects in a given document. In cognitive science, comprehension is often characterized as the construction of a mental model that represents the objects and semantic relations in a text *[15]*. Thus, throughout the reading process we want to assist readers in the construction of their mental models.

Perspective-based reading provides readers assistance in form of operational scenarios (in short: scenarios). A *scenario* of a perspective is an algorithmic description of activities and questions from which to build an abstraction of the inspected document and analyse it. It is developed based on knowledge about the environment in which the reading process is applied: roles in the software development process and defect classes, see Figure 1.

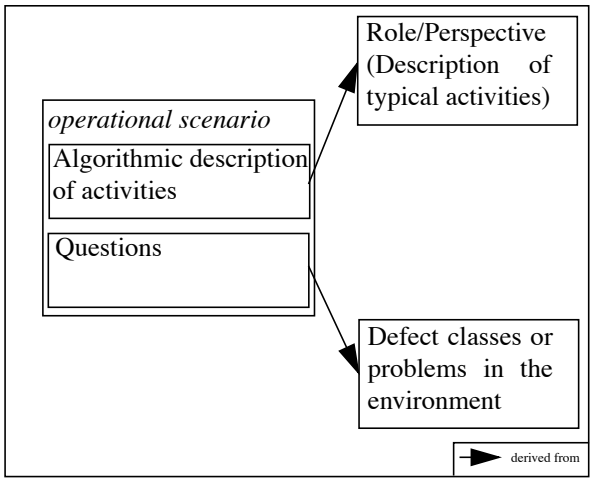

Figure 1. Description of the Model

An *activity* of a scenario is a description on how to build an abstraction of the inspected document. An activity should be typical for a particular role within the software development process. The role determines the perspective from which the reader is to inspect the document, typically a customer or consumer of the document. For example, building test cases is a typical activity for a tester (role). Thus, the same activity needs to be performed by a reader reading from the perspective of a tester.
A *question* is an interrogation of the reader about the activity, i.e. the process of building the abstraction or the result of the activity. The questions are derived from defect classes or problems that are typical for the product or for the environment. The questions on the scenario should not be compared to the tick-list of a checklist. For example, compare a typical question of a checklist

"Are all items sufficiently and unambiguously described?"

with a typical question of a scenario for a tester answered while performing an activity:

"Do you have all the information necessary to identify the item being tested and to identify your test criteria? Can you make up reasonable test cases for each item based upon the criteria?

Using a scenario of the technique Perspective-based reading provides a means for the reader to structure and reduce the information of the inspected document by building an abstraction. We hypothesize that this will facilitate the construction of a mental model in the course of the reading. A better mental model will lead the reader to a better and focused understanding of the inspected document which helps him detecting more defects than before.

In the original experiment, the documents were read from three different perspectives: A tester perspective (T), a user perspective (U), and a designer perspective (D). For each of these perspectives one scenario was available which we used in our replication. This allows us in the following to treat the terms "perspective" and "scenario" as synonyms. However, the reader must keep in mind that there are other situations in which two scenarios for one perspective may be used. Appendix A contains a complete example of the scenario for reading from the perspective of a tester.

2.2 Goals, Hypotheses, and Variables

The overall goal of our replication is to examine whether PBR as opposed to Ad-hoc reading can help improve defect detection effectiveness. To document our goal, and the derived hypotheses and variables, we use the GQM approach *[4, 9, 10, 12]*. The GQM approach has been developed by Basili and Rombach to support the goal-oriented derivation and interpretation and interpretation of measurement in software engineering. It has successfully

been applied in industrial measurement programs *[14, 29]*, but also to document experimental goals and the hypotheses and variables derived *[28]*. We found that the usage of techniques of the GQM approach like GQM plans (we use the terminology introduced in *[12])* beneficial to support the documentation of experiments. This makes further replications easier. Our GQM plans document experimental goals, hypotheses, variables, and the analysis performed. In this section goals, hypotheses, and variables are organized and structured through GQM plans. Additional elements of the GQM plans are the questions, which state informational needs in an informal way. To answer the questions, measures are derived. They represent the experimental variables. Section 2.6 shows how a GQM plan can be extended to plan data analysis. In the following, parts of GQM plans will be emphasized by usage of a different font.

We state the overall experimental goal using the GQM goal template:

Analyze PBR and Ad-hoc reading techniques
for the purpose of their evaluation
with respect to their effectiveness
from the viewpoint of the researcher
in the context of the Software Engineering (SE) lecture at the University of Kaiserslautern (UKL).

In our replication, we investigate four different facets of the quality focus effectiveness (introduced in the template by the formulation *with respect to*). Each of the facets of effectiveness relies on a different effectiveness model. This leads us to define a different GQM goal for each facet. As the viewpoint and the context do not change with respect to the experimental goal, they will be left out for repetition reasons:

Goal 1:
Analyze PBR and Ad-hoc reading techniques
for the purpose of their evaluation
with respect to their effectiveness for individuals ...

Goal 2:
Analyze PBR and Ad-hoc reading techniques
for the purpose of their evaluation
with respect to their effectiveness for teams ...

Goal 3:
Analyze PBR perspectives
for the purpose of their evaluation
with respect to their effectiveness to detect different defect classes ...

Goal 4:
Analyze PBR perspectives and their combinations
for the purpose of their evaluation
with respect to their effectiveness to detect differing defects ...

Each of the four goals leads to different hypotheses being tested within the experiment. Most of the hypotheses have been tested in the original experiment, but two of them are new. We express our hypotheses as experimental hypotheses[1]. As the hypotheses are derived from the same goals as are the measures and models used to test them, they are seen as part of the GQM plan. GQM plans support the structuring in dependent and independent variables through partitioning their elements into two different categories called Quality Focus and Process Definition (for details see *[12]*). The Quality Focus category contains the model definition for effectiveness, the questions (denoted Qn) stating the informational need, the hypotheses associated with the model, and the measures representing the dependent variables. Measures are denoted as Mn in the plan. We use an abbreviation for measures (or their data) in the text, which is denoted in brackets after each measure in the plan. The Process Definition category contains questions and measures concerning the independent variables.

An excerpt of the GQM plan is given for Goal 1:

Goal 1:
Analyze PBR and Ad-hoc reading techniques
for the purpose of their evaluation
with respect to their effectiveness for individuals....

Quality Focus
Model for effectiveness for individuals:
Mean defect detection rate ($\overline{DDR}$) for all individuals, with DDR = number of defects found by individual / total number of defects in the document.

Q1. Which is the mean defect detection rate of the individuals?

H1: Individuals applying PBR perform better than individuals using Ad-hoc reading with respect to their mean defect detection rate.

H2: Individuals applying each PBR perspective respectively perform better than individuals applying Ad-hoc reading (no perspective) with respect to their mean defect detection rate.

H3: There is no difference between individuals reading the different documents with respect to their mean defect detection rate.

H4: There is no difference between subjects applying PBR on one document and those applying Ad-hoc on the other document and vice versa with respect to their mean defect detection rate. (Interaction effect)

1. We use the term experimental hypotheses to denote our expected findings.

H5: The experience of the subjects has no influence on their mean defect detection rate.

M1.1 Number of defects found by each subject (NRDEFFOUND)
M1.2 Total number of defects in each document (NRDEFDOC)

Process Definition
Q2. Which document was used?
M2. Document used by subject (DOC)
Q3. Which technique was applied?
M3. Reading technique used by subject (RTECH)
Q4. Which perspective was applied?
M4. Perspective used by subject (PERSP)
Q5. How much experience did the subjects have with inspections?
M5. Experience of subject (EXP)
Q6. How well did the subjects know the english language?
M6. English knowledge of subject (ENGL)
Q7. How well did the subjects understand the technique?
M7. Understanding of technique of subject (UNDERST)
Q8. How well were the subjects motivated?
M8. Motivation of subject (MOTIV)
Q9. How well did the subjects follow the reading scenario?
M9. Following of guidelines of subject (FOLLOW)
Q10. Did the subjects have enough time to complete their work?
M10.1. Enough time for subject (TIME)
M10.2. Percentage of document finished (PERCENT)

The Quality Focus category of the GQM plan defines the dependent variable (effectiveness for individuals) and contains the hypotheses concerning this variable and the main independent variables (DOC, RTECH, PERSP, EXP). The Process Definition Category contains questions concerning the independent variables (Q2-Q5), with Q2, Q3 and Q4 asking for controlled variables. Question Q5 asks for a variable which is considered to have an influence (by hypothesis H5), but cannot be controlled (EXP). This is an uncontrolled independent variable which has to be collected. Questions Q6 to Q10 consider information concerning the process conformance of the subjects. This information is used to evaluate the validity of the collected data. For example, if the subject does not know the english language, their defect detection data will not give much insight into using a scenario in english language while inspecting a requirements document in english language.

The GQM plan for Goal 2 reuses the whole Process Definition category of the GQM plan for Goal 1. A new model for effectiveness is defined, and a new controlled variable appears: the technique to build teams.

Goal 2:
Analyze PBR and Ad-hoc reading techniques
for the purpose of their evaluation
with respect to their effectiveness for teams ...

Quality Focus:
Model for effectiveness for teams
Mean defect detection rate ($\overline{DDR}$) for the teams, with DDR = number of defects found by team / number of defects in the document. A team is a group of at most three individuals.

Q11. Which was the mean defect detection rate of teams?

H6: Nominal PBR teams detect more defects than nominal Ad-hoc teams, i.e. they have a higher mean defect detection rate.

H7: (New) The defect detection rate of nominal teams is lower than the defect detection rate of the real teams.

M1.1: Number of defects found per team (NRDEFTEAM)
M1.2: Number of defects in document (NRDEFDOC)

Process Definition
Q12. Which team building technique was used?
M12: Technique to build teams (nominal or real) (TTECH)

An new controlled independent variable (TTECH) is needed in order to investigate hypothesis H6. TTECH defines the used team building procedure. The variable captures the way in which teams were built: by simulation or by actual conduction of team meetings. The simulation of teams is described further in Section 2.6.

The GQM plan for Goal 3 is new to the replication. It had not been investigated in the original experiment. The plan for Goal 3 reuses the Process Definition category of Goal 1 except Q3, i.e. the question considering the applied technique. Q3 is not needed, because only the perspectives of PBR are considered here. In the Quality Focus category, a new model for effectiveness is derived together with a hypothesis to investigate it.

Goal 3:
Analyze the PBR perspectives
for the purpose of their evaluation
with respect to their effectiveness to detect different defect classes ....

Quality Focus:
Model for effectiveness to detect defect classes:
Mean number of defects per class found by the perspective / Total number of defects per class in the document (For an extensive discussion of this model see Section 2.6.)

Q13. Which was the defect detection rate of individuals for the different PBR perspectives across different defect classes?

H8: (New) For any PBR perspective, there is a difference between the mean defect detection rates of different defect classes compared to the total mean defect detection rate.

M13.1: Number of defects of each class detected by the perspective (NRDEFCLASSPERSP)
M13.2: Total number of defects of each class in the document (NRDEFCLASSDOC)

Goal 4 reuses the Process Definition category of GQM plan for Goal 3. A new Quality Focus category is derived.

Goal 4:
Analyze PBR perspectives and their combinations
for the purpose of their evaluation
with respect to their effectiveness to detect differing defects
...

Quality Focus
Effectiveness to detect differing defects:
Number of commonly detected defects by perspectives for the following groups: the three perspectives respectively, three combinations of two perspectives, and all three perspectives together

Q14. How many defects were found by different combinations of perspectives?

H9: The overlap of commonly detected defects among perspectives is low.

M14.1: Number of defects found by different combinations of perspectives (NRDEFCOMB)
M14.2: Total number of defects found (NRDEFFOUND)
M14.3: Reuse M1.2 (NRDEFDOC)

From the overall goal to investigate effectiveness of two reading techniques, we derived four sub-goals, based on the underlying models for effectiveness. GQM plans were beneficial to structure hypotheses and variables. The interrelationships of goals, hypotheses and variables are naturally documented by the plan structure. This structuring is used in Section 2.6 where the GQM plan for Goal 1 is extended to plan and document our data analysis.

2.3 Experimental design

The controlled independent variables (DOC, RTECH, PERSP, TTECH) derived in Section 2.2 set constraints for possible experimental designs. While designing the experiment based on *[7]* special attention has been paid to the following aspects:

Document choice. We only used the generic domain with the documents ATM (a specification for an automated teller machine network) and PG (a specification for a parking garage control system) because the documents of the NASA domain require a deep understanding of flight dynamic applications which only NASA employees have.

Review sequence. Each document should be inspected with both techniques to control for differences between the documents. Subjects should read each document only once, because the knowledge about defects in the document would disturb the results the second time. The order with which the techniques are applied may also have undesirable influences on the result, because subjects starting with PBR may continue applying PBR even though they are supposed to use Ad-hoc reading. Therefore, all subjects start with Ad-hoc reading, and then apply PBR.

Training. In the experiment, we did not have enough time to provide training sessions for PBR *and* Ad-hoc reading because of lecture constraints. We decided not to give training for Ad-hoc reading, but only for PBR. In the first run conducted during Winter Semester 1995/1996, we realized that it was difficult for the students having the PBR training and PBR reading on one day because of schedule constraints. So we decided in the second run, conducted during Winter Semester 1996/1997, to provide the training on an additional day. In the following the two runs of the experiment are denoted as WS 95/96 run and WS 96/97 run.

Implementation of real team meetings. In the original experiment no real team meetings were performed. Team scores were "simulated" by using the union of defects detected by the subjects in nominal teams. This ignores the impact real team meetings and real interaction between team members could have on the effectiveness of teams. Other experiments *[30,37]* reported that the real team meetings had little influence on defect detection effectiveness; thus, ignoring real team meetings may have only little influence on the effectiveness. We decided to investigate whether simulating teams is a valid approach: The last day of the second run of the experiment, we performed real team meetings and examined whether the real teams detected more defects than nominal teams.

Considering all these constraints, the resulting design is a 2x2 fractional factorial repeated measures design (see Figure 2). The subjects apply two reading techniques (controlled independent variables RTECH and PERSP) to two requirements documents (controlled independent variable DOC). The controlled independent variable TTECH effects only the data analysis because nominal teams are out of interest during the experimental procedure.

The subjects have been assigned randomly to the groups, the perspectives, and the real teams. This has been done by assigning identification numbers randomly to subjects and using assignment tables as shown in Table 1.

Group 1	Group 2	
PG / Ad-hoc	ATM / Ad-hoc	day 1
training for PBR		day 2
ATM / PBR	PG / PBR	

WS 95/96 run

Group 1	Group 2	
PG / Ad-hoc	ATM / Ad-hoc	day 1
training for PBR		day 2
ATM / PBR	PG / PBR	day 3
PBR inspection meetings (real teams)		day 4

WS 96/97 run

Figure 2. 2x2 factorial experiment with repeated measures in block of size 2

ID	Group 1			Group 2		
	D	T	U	D	T	U
1	X					
2		X				
3			X			
4				X		
5					X	
6						X
7	X					
8		X				
...	...					

Table 1. Subject Assignment to PBR Perspectives and Groups
D=Designer, T=Tester, U=User Perspective

2.4 Subjects

The subjects of the experiment were students of the Computer Science Department at the University of Kaiserslautern, Germany, who enrolled in the basic Software Engineering course lasting a semester. This course teaches the basic Software Engineering principles. The course is supplemented by practical exercises.

For the WS 95/96 run, the students were asked if they wanted to participate in a practical exercise for inspecting requirements documents (i.e. they were not asked to participate in an experiment). This run lasted two days. 35 students participated in total; 34 were present the first day, 26 on the second day, and 25 were present on both days of the experiment. In the WS 96/97 run, the students were aware of participating in an experiment to evaluate reading techniques. This run lasted four days; in difference to the first run the second day served for an extended training session for PBR and the fourth for real team meetings. We had 38 subjects in total; 31 were present the first day, 33 on the second and third day, while 26 students participated on the first three days of the experiment. The real team meetings were attended by 16 students.

The students participating were volunteers. We did not offer any incentives for the WS 95/96 run. For the WS 96/97 run, we offered the students a certificate for the training in a new reading approach in order to motivate them to collaborate. This has proven successful, a much larger part of the students attending the course volunteered. The students did not have to attend the meeting part of the

experiment to receive the certificate. We perceive that this was the reason why quite a number did not participate in the real team meetings.

All students had their Vordiplom, an initial set of exams which students have to pass after at least two years and which includes theoretical, practical and technical Computer Science, mathematics and an elective class. The subjects reported an average motivation for attending the experiment on a subjective scale (0..5, (0 = no motivation, 5 = highly motivated), minimum 1, median 3, maximum 4), an average knowledge of English on a subjective scale (0..5, (0 = no english knowledge, 5 = fluent english), minimum 2, median 3, maximum 5), and no experience in writing requirements documents on a subjective scale (0..3, (0 = no experience, 3 = more than two requirements documents), minimum 0, median 0, maximum 1). No subject had experience in inspections

2.5 Experimental Materials

We used three different requirements documents: One for the training and two for the real experiment. The requirements documents were:

- A specification for a video rental system called the ABC Video document. It was 14 pages long and contained 16 defects. This document was used for the training.
- A specification for an automated teller machine network, called the ATM document. It was 17 pages long and contained 29 defects.
- A specification for a parking garage control system called PG document. It was 16 pages long and contained 27 defects.

The requirements specifications were structured according to the IEEE standard 830-1993 *[22]* and the different requirements were stated in natural language. Thus, the subjects did not have to learn any particular formalism for requirements specification. No defects were included in the first parts of the documents, i.e. the introduction and the general system description. The subjects were informed about this. The defects were seeded into the documents. As the documents were already used several times, we can assume that the list of known defects is almost complete.

2.6 Data Analysis Procedure

We perform the analysis for individuals (Goal 1) and the analysis for teams (Goal 2) separately following the procedure described in *[8]*. *F*or the individual analysis we perform an analysis of variance, for the team analysis we use a randomization test *[17]*. We perform a Wilcoxon signed rank test for the comparison of real and nominal teams. For the analysis according to defect classes we use a binomial test (Goal 3). As in the original experiment, we use qualitative analysis to discuss the overlap of various perspectives as in the original experiment (Goal 4). In the next subsections we describe the analysis procedures in more detail. The analysis procedure for individuals is also described by an extended GQM plan to demonstrate its usefulness for documentation and replication purposes.

Analysis Procedure for Individuals (Goal 1)

Goal 1 compares the effectiveness of the two reading techniques with respect to scores of individuals. For the analysis of individuals, we have identified two independent variables at two levels respectively: The reading technique (RTECH) with levels PBR and Ad-hoc reading and the document inspected (DOC) with levels PG and ATM.

Basili *et al*. emphasize that the design implies that the interaction between the document and the reading technique (RTECH x DOC) is totally confounded within the group effect, which means that the interaction between the document and reading technique cannot be examined separately from the effect group assignment has. According to *[7]* it is assumed that the RTECH x DOC interaction does not have an important influence on the dependent variable, because the documents are generic, i.e. they are similar enough that the other group would perform similar on the document if the same technique is used.

The GQM plan for Goal 1 can be further refined to support the planning and documentation of the analysis procedure:

Goal 1:
Analyze PBR and Ad-hoc reading techniques
for the purpose of evaluation
with respect to their effectiveness for individuals
from the viewpoint of the researcher
in the context of the SE course at UKL.

Figure 3 gives an overview of the refinement structure. It contains the goal, the upper levels of the questions, and the investigated hypotheses. The hypotheses are described as links between the dependent and the independent variables, or, stated in GQM terminology, between questions of the Quality Focus category and the Process Definition category. For each hypothesis, the planned or used testing procedure as well as the result of the test can be added in the GQM plan. We did not do this here as the results are only described later.

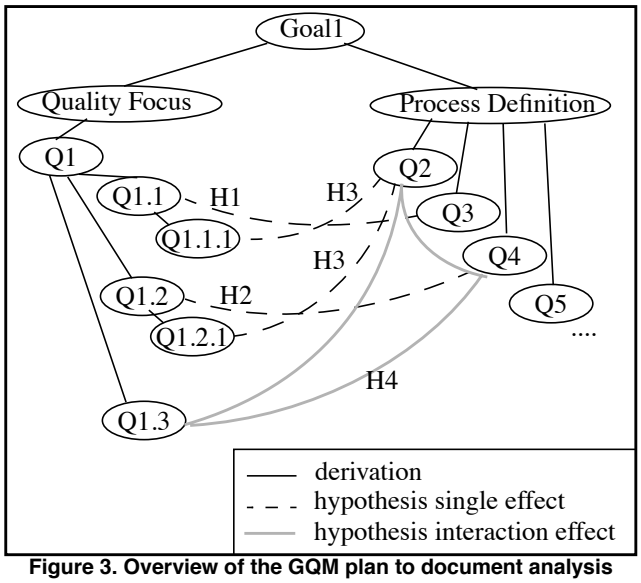

Figure 3. Overview of the GQM plan to document analysis

Quality Focus
Model for effectiveness for individuals
Mean defect detection rate ($\overline{DDR}$) for all individuals, with DDR = number of real defects found by subject / number of defects in the document.

Q1. Which is the mean defect detection rate ($\overline{DDR}$) of the individuals depending on the applied techniques, the applied perspective, and the document read?
Q1.1. Which is $\overline{DDR}$ for the applied techniques, and how does $\overline{DDR}$ differ for the applied techniques depending on the document read? *This question investigates H1(PBR vs.* Ad-hoc) *and H3 (ATM vs. PG).*

Q1.1.1. Which is $\overline{DDR}$ for PBR, and how does $\overline{DDR}$ for PBR differ depending on the documents read?
Q1.1.1.1.Which is $\overline{DDR}$ for PBR for ATM?
Q1.1.1.2.Which is $\overline{DDR}$ for PBR for PG?

Q1.1.2.Which is $\overline{DDR}$ for Ad-hoc, and how does $\overline{DDR}$ for Ad-hoc differ depending on the documents read?
Q1.1.2.1.Which is $\overline{DDR}$ for Ad-hoc for ATM?
Q1.1.2.2.Which is $\overline{DDR}$ for Ad-hoc for PG?

Q1.2. Which is $\overline{DDR}$ for the applied perspectives, and how does $\overline{DDR}$ differ for the applied perspective depending on the document read? *This question investigates H2 (user vs. tester vs. designer perspective) and H3 (ATM vs. PG).*

Q1.2.1. Which is $\overline{DDR}$ for the user perspective, and how does $\overline{DDR}$ for the user perspective differ depending on the documents read?
Q1.2.1.1.Which is $\overline{DDR}$ for the user perspective for ATM?
Q1.2.1.2.Which is $\overline{DDR}$ for the user perspective for PG?

Q1.2.2.Which is $\overline{DDR}$ for the tester perspective, and how does $\overline{DDR}$ for the tester perspective differ depending on the documents read?
Q1.2.2.1.Which is $\overline{DDR}$ for the tester perspective for ATM?
Q1.2.2.2.Which is $\overline{DDR}$ for the tester perspective for PG?

Q1.2.3.Which is $\overline{DDR}$ for the developer perspective, and how does $\overline{DDR}$ for the developer perspective differ depending on the documents read?
Q1.2.3.1.Which is $\overline{DDR}$ for the developer perspective for ATM?
Q1.2.3.2.Which is $\overline{DDR}$ for the developer perspective for PG?

Q1.3. Which is $\overline{DDR}$ for the applied technique and the document read considered in interaction, and does $\overline{DDR}$ differ for the applied techniques and the document read? *This question investigates H4 (user vs. tester vs. designer perspective in interaction with ATM vs. PG)*

Process Definition
Q2. Which document was used?
M2. DOC = (ATM, PG)
Q3. Which technique was applied?
M3. RTECH = (PBR, Ad-hoc)
Q4. Which perspective was applied?
M4. PERSP = (User, tester, Designer)
Q5. How much experience did the subjects have with inspections?
M5. EXP = (Poor, Much, Very much)
. . .

Looking at the refinement structure of Q1 (see Figure 3) we see that hypotheses H1 and H2 are only investigated in combination with hypothesis H3. Moreover, we see that mainly the effects of the controlled independent variables on the dependent variable are tested. Of course, the same investigation can be done for the uncontrolled independent variable EXP. The influence of the most important measures from the Process Definition category on the measures and models of the Quality Focus category are tested according to the stated hypotheses. For Q6 - Q10, an analogous analysis planning and documentation can be performed, but is not described here.

The GQM plan for Goal 1 and the corresponding figure show that the structure of the GQM plan naturally supports the structuring of the analysis. Therefore it showed to be very useful to document the analysis for replication purposes.

Analysis Procedure for Individuals (Perspectives)

The analysis for individuals described above examines if the technique has an influence on the defect detection rate of individual reviewers. However, it is also interesting to see whether the perspective also has an influence, i.e., if the

application of a single perspective increases the defect detection rate .

Every reviewer had to apply Ad-hoc reading and one PBR perspective to different examples. Thus, we have four groups of subjects for each document: Those applying Ad-hoc reading, and those using any PBR perspective. If we do not consider the documents separately, every reviewer is in two groups, the Ad-hoc group and a PBR perspective group. Because the same number of subjects are assigned to Ad-hoc and to PBR for each document, the design also implies that the group that contains the Ad-hoc readers is three times larger than the other groups containing the single perspectives. Therefore, we cannot just run an ANOVA like for the technique analysis, because ANOVA uses an F-test that is only reliable if the largest group in the test is no more than 1.5 times larger than the smallest group in the test or if the groups are homogeneous - and we cannot guarantee that. In order to make a valid comparison between subjects using Ad-hoc and subjects using any perspective, we build subgroups of the Ad-hoc groups. A subgroup contains subjects that applied first Ad-hoc reading on one document and then applied a specific perspective on the second document. This implies that all subjects are present in both groups of the analysis, the groups are of equal size and each reviewer serves as his or her own control object. This setting also implies that each group of the analysis contains scores of readers for both documents. This is possible, because we have already shown that the difference between the scores for the two documents is non-significant (see Table 4 for the WS 95/96 analysis and Table 7 for the WS 96/97 analysis).

We conducted an ANOVA (using an F-test) testing the null hypothesis that the perspective has no influence on the scores of individuals. We decided to increase the significance level to α=0.1 because of the low number of data points; therefore, if we obtain a p-value below our significance level of 10%, we can reject the null hypothesis and conclude that the perspective has an influence on the defect detection rate. We expect that the subjects applying PBR perspectives have higher defect detection rates than reviewers performing Ad-hoc reading.

Analysis Procedure for Nominal Teams (Goal 2)

For the analysis of nominal teams we only provide a brief overview. The problems encountered and the analysis procedure are well described in *[7]* and *[8]*. The rationale behind permutation tests is described in *[17]*.

For nominal teams, we cannot just compare the results, for example, by running an F-test because results of every subject are considered in more than one nominal team *[8]*. Thus, these teams are not statistically independent. To avoid the problem, we use a permutation test *[17]*. As a test statistic we use

$$(d_{A,PBR} - d_{A,Ad\text{-}hoc}) + (d_{B,PBR} - d_{B,Ad\text{-}hoc})$$

where $d_{A,PBR}$, $d_{B,Ad\text{-}hoc}$, $d_{A,Ad\text{-}hoc}$ and $d_{B,PBR}$ are the mean defect detection rates of all possible teams of group 1 on document A and B, and for all possible teams of group 2 on document A and B, respectively. This test statistic indicates the relative advantage of PBR compared to Ad-hoc.

The test statistic is compared with the test statistic that would be achieved if some subjects were exchanged from one group to the other, and the comparison is a direct indicator for the benefit of PBR. If no difference exists between the groups of teams applying PBR and Ad-hoc reading, and if we then switch readers from one group to the other and perform the same analysis again, any difference in the test statistics will only be due to random effects. In other words: Let us first assume that for the document A, we exchanged one reader from group 1 to group 2, i.e. from the PBR group to the Ad-hoc group and pretend that this reader did in fact perform Ad-hoc reading. At the same time, we switch one Ad-hoc reader for the same document to the PBR group (i.e. from group 2 to group 1), and let us pretend that s/he in fact performs PBR, applying no specific perspective. For the other document, we do the corresponding "swap" including the same two subjects - this guarantees that the basic assumption holds: every inspector reads the two documents applying different techniques. The effect is, that the reader from group 1 is considered part of the Ad-hoc group for document A and part of the PBR group for document B, even though he applied PBR on document A and Ad-hoc reading on document B. The other subject is in a similar but reversed situation: S/he is assumed to have applied PBR on document A and Ad-hoc reading on document B, even though this is not true. Of course, there will be a difference in the test statistic between this "diluted" groups and the "undiluted" groups where no subjects are exchanged between the groups.

This procedure is now repeated for all possible assignments of subjects to groups, i.e. all possible "swaps" of inspectors between group 1 and group 2. For all these "swaps", the test statistic is computed. The term "swap" denotes here any specific "virtual" assignment of subjects to group 1 or group 2. Then, these swaps are ranked according to the value of the test statistic in decreasing order. Then, the likelihood that the rank of the undiluted swaps, X, is among the top k places of the list, can be stated as:

$$P(X \le k) = \frac{k}{N}, \text{ where N is the numer of permutations created.}$$

As the permutation test assumes no difference between the scores of PBR teams and Ad-hoc teams, the significance value can be computed as the probability that the first undiluted swap can be found among the top k places of the

list, and this probability is exactly $P(X \leq k)$! The rank of the first undiluted Ad-hoc swap can be used as direct indicator for the significance value: In the original experiment 5% as significance level were chosen. Thus, we can reject H_0 (i.e. there is no difference between nominal PBR teams and nominal Ad-hoc teams with respect to the mean defect detection rate) if the first Ad-hoc swap in which no dilution occurs appears in the top 5% of the list. Then, we can conclude that PBR has a beneficial effect on team scores.

The reader must be aware that this is a "simulation" of teams because the members do not interact in any way. Thus meeting effects like meeting gains or meetings losses are not considered. However, during the second run of our experiment we performed some meetings which allow us to compare the results of the simulated meeting with the actual meeting.

Comparison of nominal and real Teams (Goal 2)

As we only performed real team meetings in the second run of our replication, the number of available data points is low. However, we run a Wilcoxon signed rank test *[36]* to investigate whether there is a statistically significant difference between nominal and real teams.

Analysis Procedure for Defect classes (Goal 3)

One main goal for PBR is that perspectives are focused to provide particular coverage of part of the document, and a combination of perspectives provides coverage of the entire document. It is assumed that different perspectives concentrate on the detection of defects belonging to different defect classes. To check this assumption we performed a qualitative and a quantitative analysis. The defects in the document are classified according to the following classification scheme: A (ambiguous information), E (extraneous), I (incorrect fact), O (omission) and M (miscellaneous). The qualitative analysis consists of a graph that maps the perspectives to the mean defect detection rates for each defect class.

We first consider each perspective separately. We examined if some defect classes were found more or less often than others, i.e. if some classes were "easier" to find than others. As we examined each single perspective, we consider the defects found by each perspective separately, i.e. for each perspective, we consider the mean number of defects found for each defect class. We assume that the defect class plays no role in the detection of defects, i.e. defects of a specific defect class are not detected more or less often than defects of another class. Then, the detected defects should be equally distributed among defect classes. These expected values can now be compared with the actual values, i.e. the true **(mean)** number of defects found for each class, using a statistical test. If there is a significant difference between the expected and actual values, we can conclude that the perspective finds this specific combination more or less often than expected, i.e. the assumption that the defect class plays no role in the detection of defects does not hold. For this analysis we combined the results of the two documents to achieve greater number of defects in each group.

As a statistical test we could not use a Chi-square test, because the approximation used for the computation of this test is only valid if at most 20% of the expected-values is below 5 and no expected-value is less than 1. In our case, both restrictions are violated. Thus, we decided to apply a binomial test to search for significant differences between actual and expected number of defects for each defect class. for any defect class, the probability for each defect to be detected is $p = 0.2$. Corresponding, the probability that it is not detected is $q = 1\text{-}p = 0.8$. If our assumption holds, we will have a binomial distribution for defects of each class. Thus, we expect $E_{class} = p * N_{class}$ defects to be detected (N_{class} denotes the number of defects found for *class*). Let the actual (observed) number of defects in the defect class be O_{class}. Then, the probability that the number of observed defects is X can be expressed as the binomial function

$$bin(X, p, N_{class}) = \binom{N_{class}}{X} p^X (1-p)^{N_{class}-X}$$

The probability that X differs from E at least like the observed value O does is calculated as:

$$P((X \leq E - |E - O|) \vee (X \geq E + |E - O|)) =$$

$$\sum_{i \leq E-|E-O|} bin(i, p, N_{class}) + \sum_{i \geq E+|E-O|} bin(i, p, N_{class})$$

where bin (n, p, N) is the binomial function as described above and N_{Class} the number of found defects in the considered class. This probability tells us how likely the difference between E and O is if the assumption about p holds.

The problem with the analysis is that the number of defects in some classes is very low; and the number of defects found by the perspective is even smaller. This means that in many cases significant values cannot be achieved, because the total number of defects in the document for the class is too low. Thus, we considered a significance level of $\alpha = 0.1$ an appropriate value.

We also tried to follow the procedure described in *[34]*. We made the following modifications: **I**n this case, we considered only perspectives as a whole, i.e. defects are found by a perspective if the are found by any reviewer

applying the perspective. We divided the defects in our documents into three groups, depending on how many readers with different perspectives had found the defect: hard, moderate and easy defects. A hard defect is one that was detected by only one perspective, a moderate defect was found by two perspectives, and we consider a defect as easy if all three perspectives found it. As we want to examine each single perspective, we consider the defects separately found by each perspective. Thus, a hard defect is a defect that is detected only by the perspective we consider, a moderate one is detected also by another perspective, and an easy defect is, as before, detected by all three perspectives. In addition, the detected defects are distributed according to their classes. Thus, we obtain for each perspective 15 possible combinations of groups (easy, moderate, hard) and defect classes (A, E, I, O, M). In an analogous way, we can calculate expected values which can be compared with the actual values, i.e. the true number of defects found for each combination of group and class, using the binomial test.

Experience and Overlap Analysis (Goal 1 and Goal 4)

We also examined the overlap of the different perspectives and the influence experience has on the results like in the original experiment. We focus on a qualitative rather than on quantitative data analysis as described in *[7]*.

2.7 Data Collection Plan and Procedure

The design of our experiment shows the allocation of subjects to tasks and the time schedule for the experimental tasks. The data collection procedures are designed to fit the experimental procedure and the GQM plans derived. Each measure contained in one of our GQM plans has to be collected at a precise point in time by each subject. We basically used the data collection forms of the original experiment, we only had to adjust them to our context and constraints. For example, we added a question about the subjects' knowledge of english, and the question if the subjects finished their task.

We used nine different forms. The subjects had to fill out the first form at the beginning of the experiment. They were asked about their motivation, experience, and knowledge. The second form gave them a detailed description of the defect classes. The third and fourth form supported the execution of the Ad-hoc and PBR reading. They gave guidance on how to mark the defects and asked if the subjects did finish their jobs. The fifth form was used to list the defects found by each subject. The sixth, seventh and eighth form contained the different PBR reading scenarios (see Appendix A for an example scenario), and the ninth was the debriefing questionnaire, asking the subjects about their opinion of the experiment in general, of the training they got etc.

2.8 Experimental Procedure

The experiment was run in the time slots of basic lectures on Software Engineering and its exercises. The length of each time slot was 1:30 hrs. This time included training and introduction, so the time for the reading task was even less.

The first run of the experiment was conducted as follows: The first part, i.e. the Ad-hoc reading, was performed in one lecture time slot. The second part was conducted on the following day and contained both PBR training and PBR reading in one block. This part had the double of the length of lectures, and therefore required some extra time from the students. As this caused problems to the schedules of the subjects, we decided to conduct the second run differently. We followed exactly the lecture and exercises schedule: Ad-hoc reading on one day, PBR training one the next. A day later, PBR reading was held, and the team meetings in the following week. All the time slots used were regular lecture and exercise slots, thus giving the students the possibility to attend.

3. Data Analysis

This Section presents the statistical analysis of the data of our two experimental runs as described in Section 2.6. Section 4.1 presents some descriptive statistics to get an overview of the data. Section 4.2 presents the analysis for individuals. Section 4.3 presents the analysis for nominal teams. Section 4.4 presents the comparison between nominal and real teams. Section 4.5 present the analysis for defect classes. Finally, Section 4.6 presents investigates the influence experience has on the results and the overlap of different perspectives.

3.1 Descriptive Statistics

First, we present some descriptive statistics and distributions of the data using boxplots. Boxplots provide information at a glance about center (median), spread (interquartile range), symetry, and outliers *[31]*. We provide boxplots for the independent variables reading technique and document for the two runs of our replication.

The WS 95/96 run

Figure 4 shows the boxplot for the independent variable document. The distribution of the defect detection rates are similar for both documents. The subjects detected between 4% and 45% of the defects in the documents.

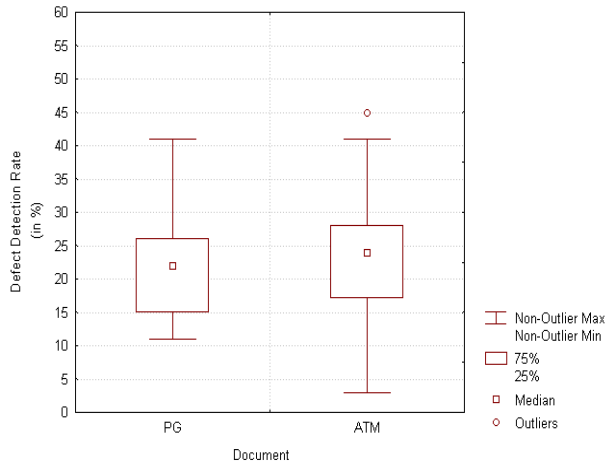

Figure 4. Box Plot of Defect Detection Rate on the two Documents

With respect to the independent variable reading technique the range of the distribution for both techniques is similar (Figure 5). Although the medians seem to be similar, the results from Perspective-based reading appear to be better regarding the interquartile range than the ones from Ad-hoc reading.

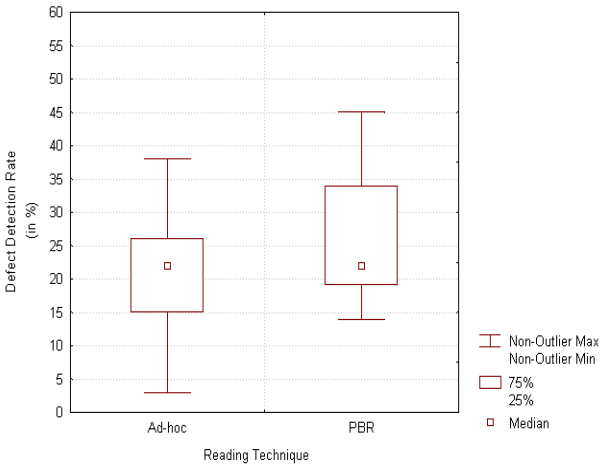

Figure 5. Box Plot of Defect Detection Rate for the Techniques

Figure 6 shows the boxplot for the independent variable perspective. The designer perspective appears to perform much better than the others, while the medians of the tester and user perspective are quite close to Ad-hoc reading ("none"). However, the two PBR perspectives appear to be better as they have smaller ranges.

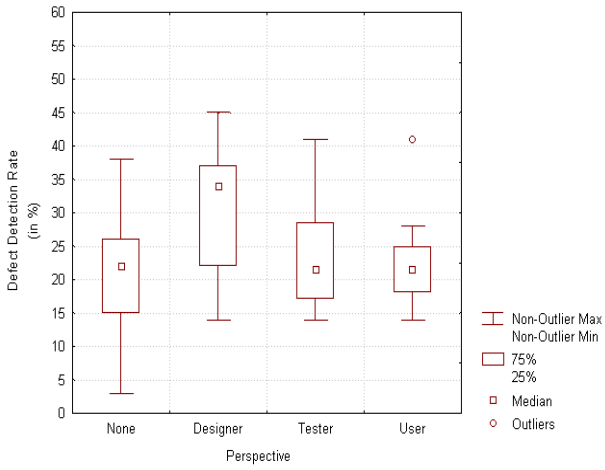

Figure 6. Box Plot of Defect Detection Rate for the Perspectives

Table 2 presents the concrete values of the mean, the standard error, and the standard deviation for the defect detection rate. Regarding the mean values, the difference between the documents is not as high as the difference between the reading techniques; even the difference between the perspectives is higher.

	Mean	Standard Error	Standard Deviation
ATM	24.440	1.949	9.747
PG	22.881	1.709	8.543
Ad-hoc	21.400	1.557	7.783
PBR	25.922	1.980	9.901
Designer	30.605	3.649	10.948
Tester	23.611	3.277	9.270
User	22.963	2.934	8.298

Table 2. Mean, Std. Deviation and Std. Error for the Defect Detection Rate

The WS 96/97 run

Figure 7, Figure 8 and Figure 9 show the results for the second run of our replication. They appear to be consistent with the results of the first run with respect to the distribution. However, the overall performance of subjects was not as good in detecting defects as in the first run, resulting generally in a lower defect detection rate.

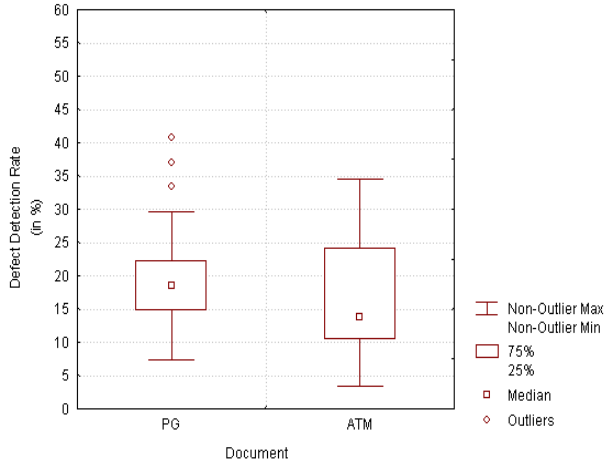

Figure 7. Box Plot of Defect Detection Rates for the two documents

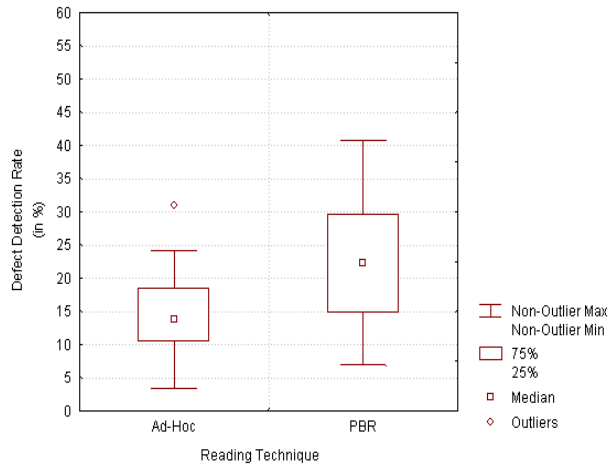

Figure 8. Box Plot of Defect Detection Rate for the Techniques

Figure 9 shows some small deviation compared with the first run. The designer perspective is not better than the other perspectives, and the difference between Ad-hoc reading and the PBR perspectives becomes more obvious.

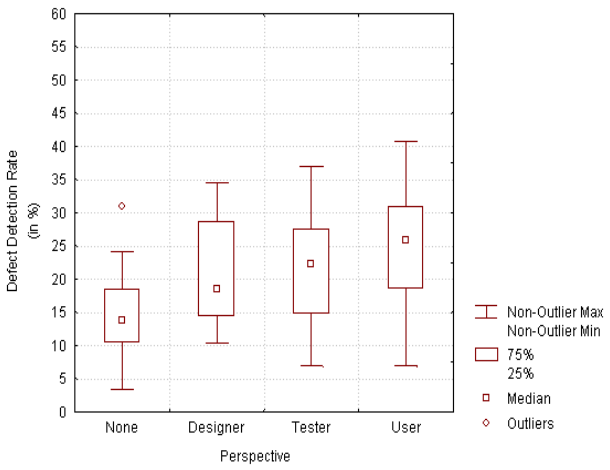

Figure 9. Box Plot of Defect Detection Rate for the Perspectives

Table 3 presents the values of the mean, the standard error, and the standard deviation for the defect detection rate. In this run, the difference in the mean defect detection rate between the reading techniques is much higher than the difference between different documents**,** while the difference between the perspectives is smaller.

	Mean	Standard Error	Standard Deviation
ATM	17.109	1.819	9.277
PG	19.943	1.632	8.320
Ad-hoc	14.589	1.236	6.301
PBR	22.463	1.835	9.355
Designer	20.961	3.114	8.808
Tester	21.371	3.203	9.609
User	24.890	3.384	10.150

Table 3. Mean, Std. Deviation and Std. Error for the Defect Detection Rate

3.2 Analysis for Individuals

Before performing the analysis as described in Section 2.6 we verified whether the assumptions described in *[7]* hold with our data set. Thus, we performed a Shapiro-Wilk W test *[35]* to make sure that each level of the independent variable is normally distributed. We also assured that the number of subjects in the largest group is no more 1.5

greater than in the smallest group. Thus, the analysis procedure is valid to use.

The WS 95/96 run

Table 4 shows that our analysis did not reveal statistical significance for both, the document effect ($p < 0.496$) and the RTECH × DOC effect ($p < 0.495$). This is what we expected (Hypotheses 1 and Hypotheses 4). The documents are similar enough that subjects have similar scores on the different documents. The reading technique also does not have a statistically significant effect on the result with respect to the chosen level of significance of 0.05. Thus we cannot reject the null hypotheses, that there is no difference between PBR and Ad-hoc reading. However, the p-value allows us to argue that a Type I error, i.e. reject the null hypothesis when it is in fact true *[36]*, only appears in 7.7 out of 100 cases. From a practical point of view this can be regarded as significant. Although the different documents suppose to be similar we performed a one-way Analysis of Variance for each document separately. We also did not get statistically significant results for the technique effect, but we found a lower p-value in the case of the PG document than the ATM document. This is explained by the fact that (1) the differences in the mean between PBR and Ad-hoc are bigger for the PG document than for the ATM document (Figure 10) and (2) the standard deviation for the PG document is smaller than for the ATM documents as seen in the last subsection.

	df Effect	MS Effect	df Error	MS Error	F	p-value
Document	1	37.94	46	80.69	0.47	.4963
Technique	1	262.9	46	80.69	3.25	.0776
Doc x Tech	1	38.22	46	80.68	0.47	.4948

Table 4. ANOVA Summary of all Effects for RTECH x DOC Interaction

Table 5 and Table 6 show the ANOVA result for the documents examined separately. It is interesting to see that, even though in both cases significant p-values were not achieved, the p-value for the ATM document is about ten times higher than for the PG document. Considering the lower number of data points for this analysis, one could raise the significance level to α=0.10; then, the result for the PG document would be significant.

	df Effect	MS Effect	df Error	MS Error	F	p-value
Technique	1	50.32	29	96.94	0.51	.4785

Table 5. ANOVA Technique Effect for ATM Document

	df Effect	MS Effect	df Error	MS Error	F	p-value
Technique	1	250.80	23	64.42	3.892	.0606

Table 6. ANOVA Technique Effect for PG Document

Figure 10 shows the mean defect detection rates of individuals. The scores for the two documents are very close to each other, again confirming that there is only a small difference between the documents.

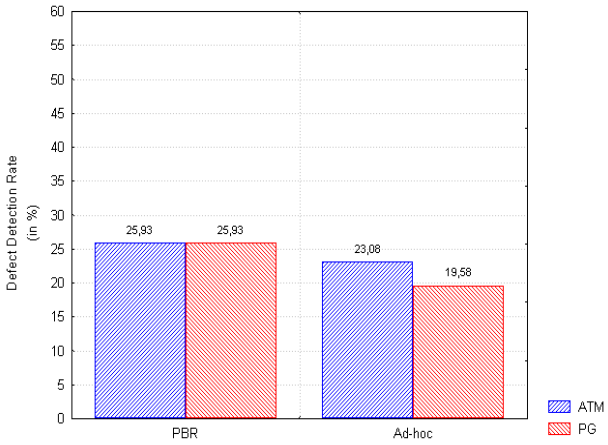

Figure 10. Mean Defect Detection Rates

The WS 96/97 run

Table 7 shows that our analysis failed to reveal statistical significance for both the document effect ($p < 0.206$) and the RTECH × DOC effect ($p < 0.577$). However, for the WS 96/97 run, the reading technique has shown to have a significant effect on the result, as the p-value is much smaller than the significance level of 0.05. In this case the differences are large enough to be of statistical significance. We also performed a One-way Analysis of Variance for this experimental run. We only find statistical significant results for the PG document, where the p-value is much smaller than the significance level. Thus, the subject found

significantly more defects with PBR than with Ad-hoc reading. For the ATM document, the p-value is very close to the significance level.Table 8 and Table 9 show the ANOVA

	df Effect	MS Effect	df Error	MS Error	F	p-value
Document	1	104.4	48	63.67	1.640	.2064
Technique*	1	808.0	48	63.67	12.65	.0008
Doc x Tech	1	20.08	48	63.67	0.315	.5770

Table 7. ANOVA Summary of all Effects for RTECH x DOC Interaction

results for the separate documents. This time, both documents have low p-values; the result for the PG document is even significant at the 0.05 level. If the significance level was raised to α=0.10 due to the lower number of data points, both results would be significant. Again, it is interesting to see that in this run the p-value for the ATM document is about 20 times higher than for the PG document.

	df Effect	MS Effect	df Error	MS Error	F	p-value
Technique	1	285.8	24	77.74	3.676	.0671

Table 8. ANOVA Technique Effect for ATM Document

	df Effect	MS Effect	df Error	MS Error	F	p-value
Technique*	1	540.25	24	49.5	10.89	.0030

Table 9. ANOVA Technique Effect for PG Document

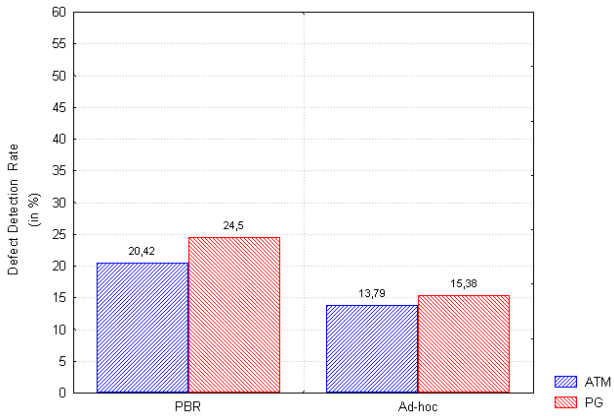

Figure 11. Mean Defect Detection Rates

3.3 Analysis for Perspectives

We also investigated whether each perspective separately performed significantly better than Ad-hoc reading (Hypothesis H2). To conduct the test described in Section 2.6, we examined carefully if the assumptions for the test were fulfilled: the independent variables (perspective, document) are measured on nominal scale, the dependent variable (the defect detection rate) on a ratio scale. The dependent variable is normally distributed for each level of the independent variables (document and perspective); we confirmed this with a Shapiro-Wilk W test ([35]). We cannot guarantee that our groups are homogeneous, but the test is robust against violations of homogeneity if the number of subjects in the largest group is no more than 1.5 times greater than in the smallest; and we designed our analysis so that all groups have equal size. A threat to the validity of this test is that our subjects are volunteers and are not obtained through random sampling.

The WS 95/96 run

The analysis, summarized in the Table 10, revealed a statistical significance only for the comparison of the Designer perspective against the Ad-hoc readers (D vs. N). Figure 12 illustrates that the user and tester perspectives have smaller defect detection rates than the designers for the ATM document; this is reflected by the corresponding p-values. The mean defect detection rates are similar for the two documents, so this shows again that the documents are similar and that their scores can be used together for analysis. It is interesting to see that the tendency observed in Section 3.1 is confirmed here, in Figure 12 and Table 10: the Designer perspective performs best of all perspectives for this run (and thus has a low p-value), while the user perspective is very close to Ad-hoc reading (also reflected in a high p-value). ((We assume that the students did not follow the instructions for the user and tester perspective

enough, because they did not have enough experience to understand the perspective))

	number of subjects	df Effect	MS Effect	df Error	MS Error	F	p-value
D vs. N*	9	1	355.5	16	63.26	5.620	.0306
U vs. N	8	1	6.250	14	66.36	.0941	.7634
T vs. N	8	1	115.5	14	99.81	1.157	.3001

Table 10Perspectives Effect for Both Documents

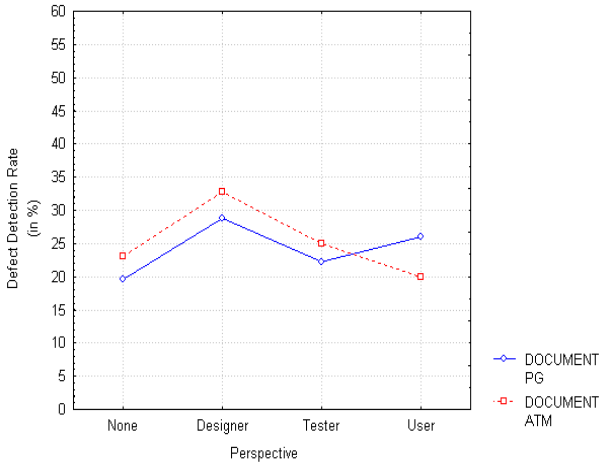

Figure 12. Mean Scores for Perspectives on Documents

The WS 96/97 run

In the WS 96/97 run, we find significant results (see Table 11) for the comparison of the Designer perspective against the Ad-hoc readers (D vs. N) and User against Ad-hoc readers (U vs. N). In this run, the User perspective is "saved" from performing worse than the other perspectives by the high performance for the PG document (see Figure 13).

	number of subjects	df Effect	MS Effect	df Error	MS Error	F	p-value
D vs. N*	8	1	246.7	14	55.27	4.464	.0530
U vs. N*	9	1	547.0	16	71.49	7.651	.0137
T vs. N	9	1	101.0	16	70.07	1.442	.2472

Table 11Perspectives Effect for Both Documents

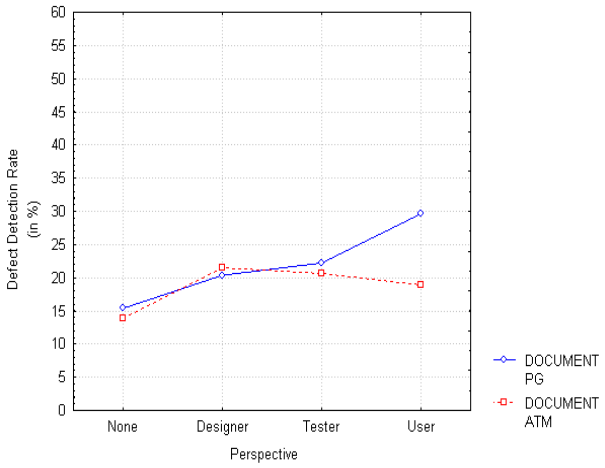

Figure 13. Mean Scores for Perspectives on Documents

We found that in almost any case the subjects did better using any perspective than reading Ad-hoc. In the WS 95/96 run the difference turned out to be statistically significant for the subjects reading from the user perspective; in the WS 96/97 run the difference turned out to be statistically significant for the subjects reading from the designer and user perspective. However, it is interesting to note that in both runs the subjects seemed to perform somewhat worse on the ATM document than on the PG document. This was also observed in the original experiment.

In conclusion, although the results of the first run failed to be statistically significant, we found an improvement of subjects using Perspective-based reading over subjects using Ad-hoc reading. In the second run the difference has been statistically significant. Thus, we can basically confirm the findings of the original experiment with respect

to the individual analysis. Table 12 presents the summary of

Hypothesis	Null hypothesis	WS 95/96	WS 96/97
H1	There is no significant difference between subjects applying PBR and Ad-hoc reading	accept H_0	reject H_0
H2	Subjects using each Perspective respectively does not perform significantly better than subjects using Ad-hoc reading	accept H_0 for user and tester, reject for designer	accept H_0 for tester, reject for user, designer
H3	There is no significant difference among subjects reading the ATM document and the PG document	accept H_0	accept H_0
H4	There is no significant difference between the independent variables reading technique and document	accept H_0	accept H_0

Table 12 Summary of findings

our findings with respect to the individual analysis and the hypotheses we stated in Section 2.2.

3.4 Analysis for Nominal Teams

The WS 95/96 run

In our WS 95/96 run, we had 35 subjects in total. Unfortunately, because the test statistic assumes that any subject has read both documents, we could only use those subjects for this analysis that had participated on both days. This results in 25 subjects whose data can be used for the test. The number of possible swaps is then $\binom{25}{12} = 5200300$, the number of possibilities of grouping 12 subjects out of 25 into one group (PBR or Ad-hoc reading) and 13 into the other. Therefore, we had to change the program that was offered in the lab package of the experiment *[8]* because this program generated one line of output for each permutation and their procedure for computing $\binom{25}{12}$ did not work properly with the high numbers that occurred. Our version only calculates the necessary statistics, i.e. the mean defect detection rate for all possible teams of the undiluted teams and the significance value for PBR teams. Because of the higher number of subjects the program tends to run a very long time. Thus, we added the possibility that the program could write down results at certain states to be able to recover and resume from the last savepoint in case of a system breakdown or reboot.

As Table 13 shows, the number of group permutations is 5 200 300, and the rank of the first undiluted Ad-hoc swap was 23 417; thus we have a significance value of 0.0045, and we can reject the null hypothesis that there is no difference between Ad-hoc and PBR teams. Figure 14 shows that PBR teams detected 22% more defects than Ad-hoc teams on average; and as our significance value is below the significance level of 0.05, we can conclude that nominal PBR teams perform actually better than nominal Ad-hoc teams.

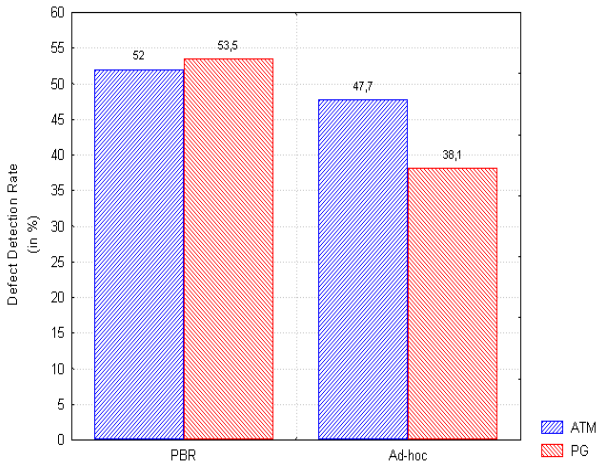

Figure 14. Mean Defect Detection Scores for all Possible Teams

	Number of Permutations created	Rank of undiluted swap	p-value
WS 95/96	5200300	23417	0.0045

Table 13. Team Analysis of the WS 95/96 run

The WS 96/97 run

In the WS 95/96 run, we had 26 subjects who participated on both days of the experiment. The number of possible swaps is then $\binom{26}{13} = 10400600$, the number of possibilities of grouping 13 subjects out of 26 into one group (PBR or Ad-hoc reading) and 13 into the other. We used our modified the program to calculate the necessary statistics.

As Table 14 shows, the number of group permutations is 10 400 600 and the rank of the first undiluted Ad-hoc swap was 3438; thus we have a significance value of 0.000331, and we can reject the null hypothesis. Figure 15 shows that PBR teams detected 48% more defects than Ad-hoc teams on average; and as our significance value is below the

significance level of 0.05, we can conclude that PBR teams perform actually better than Ad-hoc teams

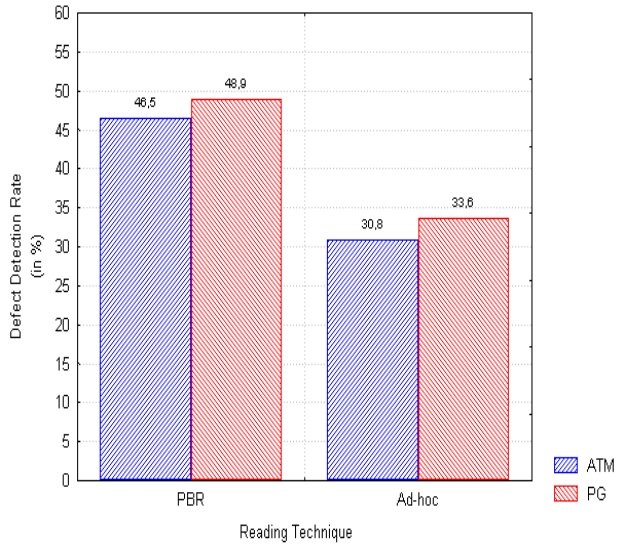

Figure 15. Mean Defect Detection Scores for all Possible Teams

	Number of Permutations created	Rank of undiluted swap	p-value
WS 96/97	10400600	3438	0.000331

Table 14. Team Analysis of the WS96/97 run

In conclusion, we have calculated a much higher number of permutations than in the original experiment. Thus, the power of the test is correspondingly higher than in the original one. With both runs of our replication, we basically confirm the results of the original experiment. Table 15

Hypothesis	Null hypothesis	WS 95/96	WS 96/97
H6	There is no significant difference in nominal PBR teams and nominal Ad-hoc teams	reject H_0	reject H_0

Table 15 Summary of findings

presents the summary of our findings with respect to nominal team results and the hypothesis we stated in Section 2.2.

3.5 Comparison of Nominal and Real Teams

The WS 96/97 run

We wanted to know how accurate the results of building nominal teams is in comparison with real teams. Thus, we decided after the first run of the replication to perform real team meetings. As all subjects have read the documents using PBR we randomly assigned three subjects in one team in which each subject has read the document from a different perspective. Unfortunately many subjects did not participate in the meetings. Thus, we changed the assignment with the available subjects resulting in two complete teams (teams with one subject of each perspective) and four incomplete teams (teams with two or with three subjects in which two have read the document from the same perspective). Table 16 shows the results (Number of logged defects in the meeting).

	Participating Perspectives	Real Team (# Logged Defects)	Nominal Team (# of Pooled Defects)	Difference between Real and Nominal Team	Meeting Gains	Meeting Losses
Complete Teams	D, T, U (ATM)	9	10	-1	0	1
	D, T, U (PG)	6	9	-3	2	5
Incomplete Teams	D, T (ATM)	14	17	-3	1	4
	D, T (ATM)	12	5	+7	9	2
	D, D, T (ATM)	10	9	+1	6	5
	D, D, T (PG)	11	15	-4	0	4

Table 16. Comparison of real and nominal team meetings

For each real team we considered the results of the comparable nominal teams, the difference between nominal and real team, the meeting gains, and meeting losses. We found no statistically significant differences between nominal and real teams, i.e. we could not reject the null hypotheses that there is no significant difference between nominal and real teams using the Wilcoxon signed rank test. In two out of six cases including both complete teams the nominal teams perform better than the real teams. Unfortunately, we did not perform meetings after Ad-hoc reading. Thus, we cannot evaluate whether nominal Ad-hoc teams also performed better than real Ad-hoc teams. However, the results suggest that meetings do not provide a synergy effect: Although some defects are newly detected in the meeting, there is also a huge number of meeting losses (defects found in the preparation but not reported in the meeting). Thus, the net meeting performance (which is the difference between meeting gain and meeting loss) is negative in four out of six cases. Hence, it might be worth to follow a different inspection strategy, i.e. to collect all defect report forms that are filled out while reading and pool the results together as input for the meeting. Thus, no defects are lost and the number of meeting losses may decrease.

Because we wanted to have a first impression about the relationship of nominal and real teams, we did not yet look at false positives, i.e. defects reported that are no defects. Based on recent results *[26,30]*, false positives seem to be a significant factor in inspection. Moreover, meetings can be used to filter out false positives. This will be a subject of future work.

3.6 Analysis for Defect Classes

We examined five defect classes: A (ambiguous information), E (extraneous), I (incorrect fact), O (omission) and M (miscellaneous). The number of defects of each class is presented in Table 17.

We also present some figures that map the perspectives to the mean defect detection rates for each defect class. This analysis is performed both for each document separately and for the both documents together.

Document	Defect Class					
	A	E	I	O	M	Total
ATM	4	2	8	14	1	29
PG	5	1	8	11	2	27

Table 17. Number of Defects in each class for the documents

The procedure for the analysis we performed is described in Section 2.6.

The WS 95/96 run

We performed the binomial test for each perspective considering the mean number of found defects and each class without looking at how often a defect was found by another perspective (see Table 24 in Appendix B). We did not find any statistically significant result. As mentioned earlier, we cannot draw reasonable conclusions from this result because the number of defects of this class is to low. We describe in more detail some qualitative results by looking at Figure 16 and Figure 17. They show the mean defect detection for every perspective and class. Quote_*X* denotes the percentage of defects found of the class *X*. For example, Quote_A denotes the percentage for the class A (ambiguous information). The small number of defects for the classes E and M (see Table 17) makes the results for Quote_E and Quote_M less valid, especially if there is only one defect: if this defect is found by chance, it raises the percentage for the whole perspective to a high level (here, 40-50%)! However, the figures show some similarities: the designer and user perspectives find a high number of inconsistencies (class I) but only a small number of ambiguities defects of class A (class A), while the tester perspective is good in finding omissions (class O) but does not detect many inconsistencies (class I). The classes M and E cannot be fully considered, because the number of defects in these classes is too low.

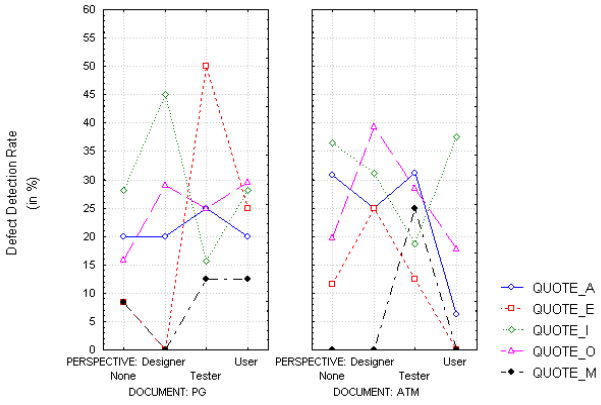

Figure 16. Mean Scores for Perspectives on Defect Classes and Document

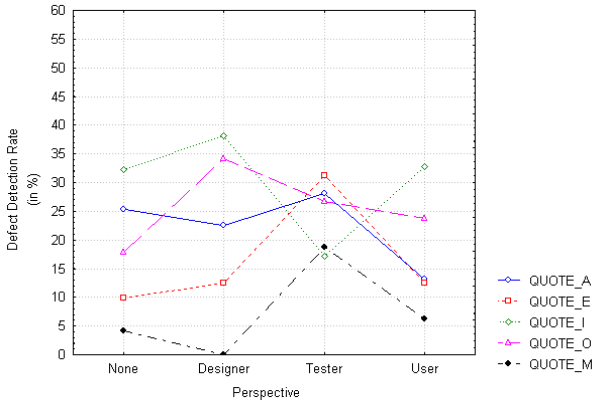

Figure 17. Mean Scores for Perspectives on Defect Classes for Combined Documents

	DDR (Total)	DDR (Class A)	DDR (Class E)	DDR (Class I)	DDR (Class O)	DDR (Class M)
None	19.85	24.85	10.29	29.78	16.16	4.41
Designer	30.61	22.22	11.11	38.89	33.62	0.00
Tester	23.61	25.00	27.78	22.22	26.19	27.78
User	22.96	13.13	12.50	32.81	23.70	6.25

Table 18. Distribution of Defect Detection Rate over Defect Classes

As we did not find statistically significant results for one perspective, we followed another analysis strategy. We investigated whether there is a difference in the results if a defect was also found by more than one perspective. We only present here the results of the statistical test we described in Section 2.6 for the tester perspective. The others can be found in Appendix B (Table 22, Table 23). An interesting aspect is that we found few significant results for the tester perspective. The column "Observed" shows the actual number of defects found for the combination of group (hard, moderate, easy) and defect class. A defect is found by a perspective, if any subject using the particular perspective has found it. "Total", "Group" and "Class" denote the number of defects found in total, for the group and for the class, respectively. Then, the "Expected" number of defects for each row is (group/total) x class. Significant differences (marked with an asterisk *) could only be found for the tester perspective, for the moderate E and I and for the easy I. It seems astonishing that there are no significant differences for the designer perspective, because Figure 16 and Figure 17 show that this perspective detects a high number of inconsistencies. Furthermore, one would expect to find a low p-value for the combinations of tester and class O, user and class I. But these figures refer to individual scores, while the analysis here only considers the union of defects found by all subjects who applied this perspective. This is the reason for the difference between the results of the two analyses.

	Class	Observed	Expected	Total	Group	Class	p-value
hard defects	A	1	0.618	34	3	7	1.000
	E	0	0.176	34	3	2	1.000
	I	0	0.618	34	3	7	0.645
	O	1	1.412	34	3	16	1.000
	M	1	0.176	34	3	2	0.169
moderate	A	3	2.059	34	9	7	0.682
	E	2	0.588	34	9	2	* 0.087
	I	0	2.059	34	9	7	0.121
	O	4	4.706	34	9	16	0.792
	M	1	0.588	34	9	2	1.000
easy defects	A	3	4.324	34	21	7	0.439
	E	0	1.235	34	21	2	0.146
	I	7	4.324	34	21	7	* 0.0499
	O	11	9.882	34	21	16	0.619
	M	0	1.235	34	21	2	0.146

Table 19. Analysis for the Tester Perspective

The WS 96/97 run

Figure 18 and Figure 19 show the mean defect detection rate for every perspective and class for the second run of our replication. We also did not get statistically significant results for the overall analysis. However, in the more detailed analysis, we found some statistically significant results for each perspective and defect class I (see Table 28 in Appendix B). Although consistent with the finding in our first run we are still limited by the small number of data points.

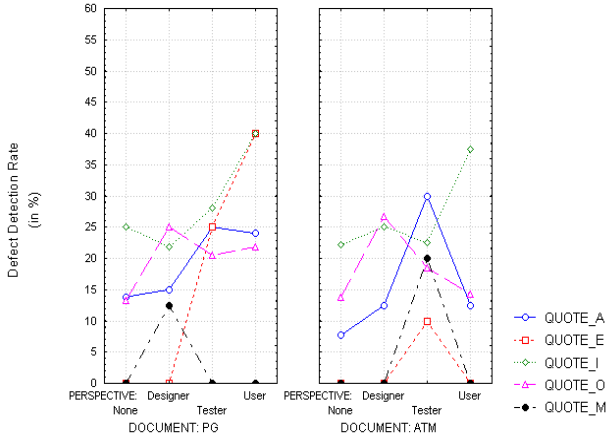

Figure 18. Mean Scores for Perspectives on Defect Classes and Document

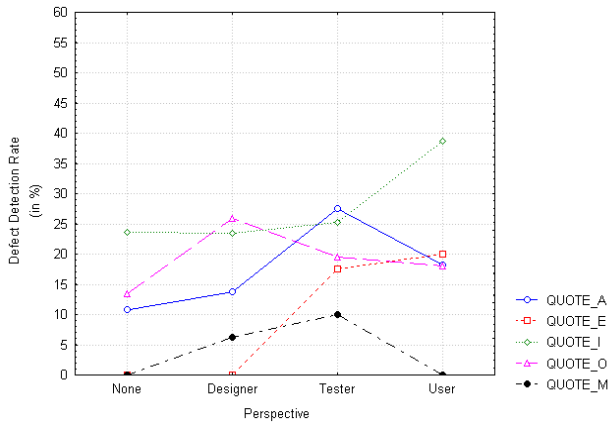

Figure 19. Mean Scores for Perspectives on Defect Classes for Combined Documents

	DDR (Total)	DDR (Class A)	DDR (Class E)	DDR (Class I)	DDR (Class O)	DDR (Class M)
None	14.59	19.77	0.00	23.56	13.51	0.00
Designer	20.96	13.75	0.00	23.44	25.89	6.25
Tester	21.37	27.78	16.67	25.00	19.41	11.11
User	24.89	18.89	22.22	38.89	18.47	0.00

Table 20. Distribution of Defect Detection Rate over Defect Classes

In conclusion, we were confronted with the low number of defects for each class and perspective. This influences the statistical procedure (use of a binomial instead of Chi-square test) as well as the results (we cannot draw many significant conclusions). Further empirical work is necessary to address this issue.

3.7 Experience and Overlap Analysis

As pointed out earlier, the analysis of the relationship between experience and defect detection rate is not interesting because our subjects all had the same experience in inspecting requirements documents (none). However, we have analyzed the overlap between different perspectives and present it within this section.

The WS 95/96 run

Figure 20 and Figure 21 show the results of the overlap analysis for the ATM and PG documents. The values we report here are mean values. In both cases, a majority of defects was found by more than one perspective, and a high number of defects is found by all three perspectives. For example, for the ATM document, 11 defects or 37.9% were found by all three perspectives, while the tester alone only found 1 defect or 3.4%. 5 defects or 17.2% were not found by any perspective. The number of defects that are found by all perspectives is particularly high; the overlap in the original experiment was not as high. One reason may be the lack of experience with requirements documents as well as the lack of experience about the various roles within the software development process. Our subjects may not have fully understood their perspective and thus fell back in Ad-hoc reading. This may explain the high overlap.

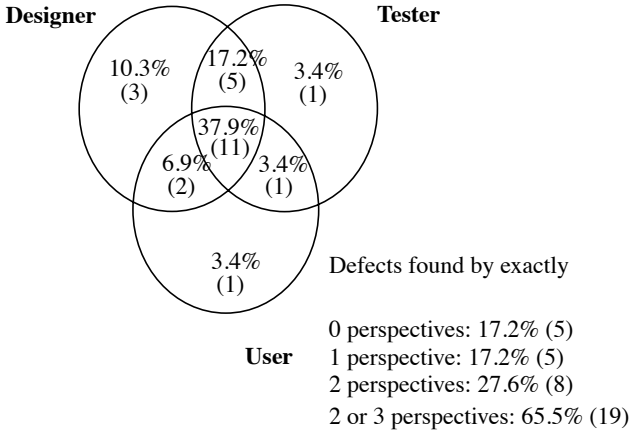

Figure 20. Overlap for ATM-Document

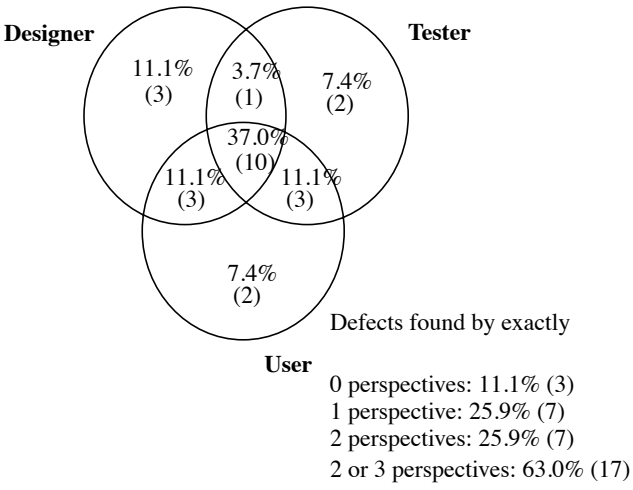

Figure 21. Overlap for PG-Document

The WS 96/97 run

Figure and Figure 23 show the results of this analysis for the ATM and PG documents for the second run of our replication. In both cases, a majority of defects was found by more than one perspective, and a high number of defects is found by all three perspectives. The subject did not find as many defects as in the first run. This may explain why the

overlap of the ATM document is not as high. However, we also observed a high overlap for the PG document.

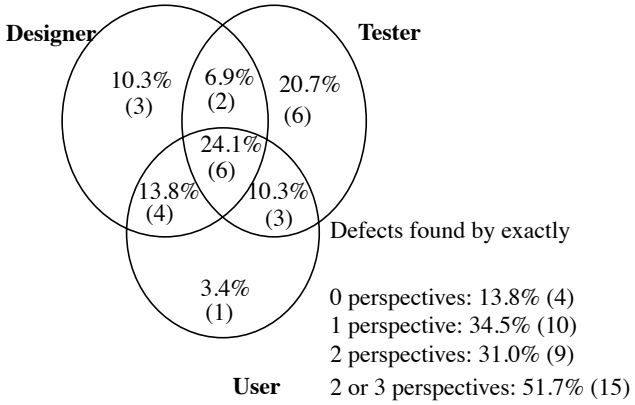

Figure 22. Overlap for ATM document

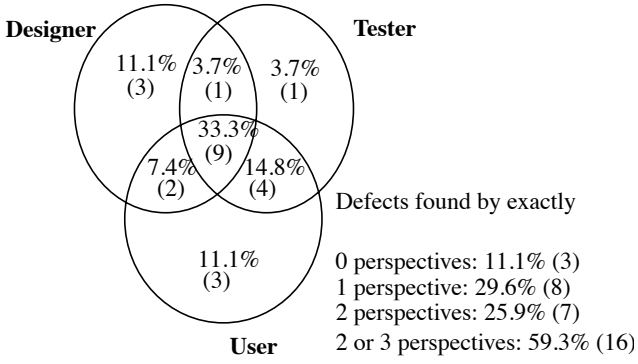

Figure 23. Overlap for PG document

4. Threats to Validity

This section discusses the various threats to validity of our experiment.

4.1 Construct Validity

Construct validity is the degree to which the independent and dependent variables accurately measure the concepts they purport to measure. In the following, we discuss construct validity with respect to effectiveness and efficiency of defect detection.

Effectiveness: There is no unique definition of effectiveness in the context of inspections. Our differentiated understanding of effectiveness concerning reading techniques as defined in Section 2.2 mainly concerns the number of real defects found in the document compared to the total number of defects in the document. This is conform with a general consensus on effectiveness of defect detection techniques as reported in various empirical studies, e.g. *[32]* and *[24]*.

Besides looking at effectiveness one can argue that time plays an important role in defect detection. Time is usually included in efficiency models for inspections. The reason why we only consider effectiveness is that our subjects only had limited time for reading due to the time constraints of our working slots. Thus, it makes no sense to investigate efficiency in our case.

4.2 Internal Validity

Threats to internal validity are unknown factors that may influence the results without our knowledge *[23]*. The following possible threats have been identified:

History. During each run of the experiment there was time in between the reading sessions in which subjects of the two groups could exchange information about defects. However, we told the subjects not to discuss the experiment nor do anything else which could influence the results. Thus, we do not consider this effect very significant.

Maturation. A maturation effect is caused by processes within the subjects that may change their behaviour. Examples are fatigue effects, loss of motivation, or learning. As the subjects in our experiment had to read only one document a day, we consider the fatigue effect less important. There may be a maturation effect from one day of the experiment to the other, causing the subjects to loose motivation on the second day (and the third day, in the WS 96/97 run) and therefore performing worse the second and/or third day. However, most of our subjects performed better the second day which is in contrast to this threat of validity. Probably the most important threat in experiments with students is learning. We consider this effect as important, because our subjects are inexperienced with requirements documents and reading techniques. We tried to overcome the learning effect with requirements documents by providing documents that have the same structure but belong to different application domains. Thus, no particular knowledge about one document could be transferred to the other and vice versa. To overcome the learning effect with PBR, we provided a training session in which the subjects used PBR on a training example. Furthermore, we discussed the training example together with the subjects.

Instrumentation. An instrumentation effect is caused by differences in the way of measuring. Instrumentation effects may result from differences in the requirements documents. We controlled for this threat by having each group inspect both documents.

Selection. Selection effects are caused by variations in human performance, e. g. knowledge. We controlled this threat by assigning randomly subjects to perspectives. Due to the random assignment we cannot assume that the

subjects are familiar with the activities performed by the perspective. This will presumably lead to an underestimation of the improvement caused by PBR and is therefore less important.

Process conformance. The subjects may not have followed the scenarios of their perspective for whatever reasons. As this effect would lead to an underestimation of the improvement caused by PBR, we do not consider this threat as grave.

Time constraints. Our subjects had only 1:30 hours to complete reading. The reason for this is that our experiment had to fit into the schedule of the course. The forms the subjects had to fill out after each reading session captured if they could finish their work, and on which page they stopped. The effect of the time restriction may be that the subjects have performed worse than without a time restriction even if they have completed their reading. This applies to both techniques, but the effect may be worse for the PBR session. The subjects may have used too much time to set up the abstraction of the inspected document (for example design) and did not have enough time to find defects. However, this leads to an underestimation of the improvement caused by PBR and is therefore less important.

Mortality. A mortality effect is caused by subjects dropping out of the experiment in a non-random fashion. Some subjects dropped out of our experiment. We cannot evaluate the effect this has on the results; but we used only data points of subjects who were present on all days of the experiment, i.e., the Ad-hoc session, the PBR training session and the PBR session. Exception is the team meeting session, where many subjects did not participate.

4.3 External Validity

Threats to external validity are problems that prevent generalizing the results of the experiment *[23]*. We identified the following possible threats:

Representative subjects. The subjects are not representative of software professionals. Our experiment was conducted with students of Computer Science who had no or little experience with requirements documents and reading techniques. This may limit the external validity of our results. This threat may only be overcome by further replication of the experiment.

Reactive arrangements. These effects are caused by the environment itself. We performed our experiment in a classroom setting. This makes it easier for subjects to concentrate on reading techniques, but of course it is not the usual working environment of the subjects.

We consider that learning and the fact that our subjects were students as the most important threats to validity in our experiment. To surmount these threats more training of the subjects and another replication of the experiment with software professionals may be appropriate.

5. Conclusions and Future Work

In the editorial of the Journal Empirical Software Engineering Victor Basili made a strong argument that researchers should build on each others' work, combine experimental results, and replicate studies under similar and differing conditions *[2]*.
In this paper we replicated an experiment originally performed at NASA/GSFC. Although we did not get statistically significant results in every case, we basically support the results and findings of the original experiment that individuals and teams perform better using PBR than an Ad-hoc approach for defect detection. *Table 15* compares the original experiment with our replication.

Perspective-based reading provides better support for individual defect detection than an Ad-hoc approach. In addition to the original experiment, we analyzed the relationship between PBR and the detection of particular defect classes. However, as the number of defects belonging to each class is low, we did not get many statistically significant results that allowed us to draw conclusions. This is also the case for the comparison of real teams and nominal teams. We found that there is no statistically significant difference between nominal and real teams. This may influence the inspection strategy that way that pooling together individual results needs to be done before the meeting. Moreover, in this experiment we found that the perspectives have a high overlap. This may be due to the fact that learning plays an important role in experiments with less experienced subjects. Thus, more empirical work is necessary in which learning as a factor can be excluded, i.e. with more training or within a case study in an industrial setting. The results of the experiment may be used to improve the PBR scenarios or to change the scenarios to improve the effectiveness of their combination. This will be topic for future work.

With respect to experimentation, we found the GQM approach a useful vehicle to document our hypotheses and analyses. We encourage other researchers to use it as it makes replication much easier. We will work on improving the maintainability and reusability of GQM plans through formalization. We intend to extend the original lab package and make a more detailed version of this paper available as a report of the International Software Engineering Research Network (ISERN) to facilitate further replication.

	Original Experiment		Our Replication	
Experimental Run	Pilot Study	1995 run	WS 95/96 run	WS 96/97 run
Subjects	12 software developers from the NASA SEL environment	14 software developers from the NASA SEL environment	25 students of University of Kaiserslautern	26 students of University of Kaiserslautern
PBR performed better than Ad hoc regarding individuals and generic documents	No statistically significant difference; PBR performed better than Ad hoc	Statistically significant difference	No statistically significant difference; PBR performed better than Ad hoc	Statistically significant difference
PBR performed better than Ad hoc regarding individuals and NASA documents	No statistically significant difference	No statistically significant difference	Not useful to use the NASA documents in our environment	Not useful to use the NASA documents in our environment
PBR performed better than Ad hoc regarding teams and generic documents	No statistically significant difference	Statistically significant difference	Statistically significant difference	Statistically significant difference
PBR performed better than Ad hoc regarding teams and NASA documents.	No statistically significant difference	Statistically significant difference	Not useful to use the NASA documents in our environment	Not useful to use the NASA documents in our environment
Influence of experience on results	Little influence of experience on results		All subjects were inexperienced. Thus, an analysis is not useful	
Analysis of real team meetings	No real team meetings performed	No real team meetings performed	No real team meetings performed	Nominal teams performed in four out of six cases better than real teams
Analysis according to different defect classes	Not analyzed		Few statistically significant results	

Table 21. Comparison of the original experiment and our replication

Acknowledgments

We thank Jean-Marc DeBaud for his comments in reviewing this paper. We thank Dieter Rombach for allowing us to perform the experiment. We thank the students of the University of Kaiserslautern for their participation in this experiment. We also thank Forrest Shull for answering our questions with respect to the analysis and we thank the Empirical Software Engineering Group at the University of Maryland for providing such an excellent laboratory manual.

References

[1] F. A. Ackerman, L. S. Buchwald, and F. H. Lewski, Software Inspections: An Effective Verification Process, IEEE Software, Vol. 6, No. 3, May 1989, pp. 31-36.

[2] V. R. Basili, Editorial, Empirical Software Engineering: An International Journal, vol. 1, no. 2, 1996.

[3] V. R. Basili,The Experience Factory and its Relation to Other Quality Approaches, In Marvin Zelkowitz, editor, Advances in Computers, Vol. 41, pp.65-82, Academic Press 1995.

[4] V. R. Basili, Applying the GQM paradigm in the experience factory. Presented at the 10th Annual CSR Workshop in Amsterdam, October 1993, to appear in a book entitled Software Quality assurance: A worldwide perspective, Chapman and Hall.

[5] V. R. Basili, Software Modeling and Measurement: The Goal/Question/Metric Paradigm, Technical Report CS-TR-2956, Departement of Computer Science, University of Maryland, College Park, MD, 20742, September 1992.

[6] V. R. Basili, G. Caldiera, H. D. Rombach, The Experience Factory, Encyclopedia of Software Engineering, Vol.1, pp. 469-476, John Wiley & Sons, 1994.

[7] V. R. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Sørumgård, M. V. Zelkowitz, The Empirical Investigation of Perspective-Based Reading, Empirical Software Engineering: An International Journal, vol. 1,no. 2, pp. 133-164, 1996.

[8] V. R. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Sørumgård, M. V. Zelkowitz, Lab package for the Empirical Investigatin of Perspective-based Reading, available on the WWW: http://www.cs.umd.edu/projects/SoftEng/ESEG/manual/pbr_package/manual.html

[9] V. R. Basili, and H. D. Rombach, A methodolgy for collecting valid Software engineering data. IEEE Transactions on Software Engineering, 10, (6), 1984.

[10] V. R. Basili, H. D. Rombach, The TAME Project: Towards Improvement-Oriented Software Environments, IEEE Transactions on Software Engineering, Vol. 14, No. 6, pp. 758-773, June 1988.

[11] V. R. Basili, R. Selby, D. Hutchens, Experimentation in Software Engineering, IEEE Transactions on Software Engineering, vol. SE-12, no. 7, July 1986.

[12] L. Briand, C. Differding and H. D. Rombach, Practical Guidelines for measurement-based improvement. To appear in Software Process Improvement and Practice, Vol.2, Issue 4, 1997. Also available as Technical Report of the International Software Engineering Research Network (ISERN-96-05).

[13] A. Brooks, J. Daly, J. Miller, M. Roper, M. Wood Replication‘s Role in Experimental Computer Science, Technical Report EFOCS-5-94, 1994.

[14] M. Daskalantonakis, A Practical View of Software Measurement and Implementation Experience within Motorola. IEEE Transactions on Software Engineering, 18, (11), 1992.
[15] T. A. van Dijk, W. Kintsch, Strategies of discourse comprehension, Academic Press, Orlando, 1983.
[16] R. H. Dunn, Software Defect Removal, McGrawHill Book Company, 1984.
[17] E. S. Edgington, Randomization Tests, New York, NY, Marcel Dekker Inc., 1987.
[18] M. Fagan, Design and code inspections to reduce errors in program development, IBM Systems Journal, vol. 15, no. 3, pp. 219-248, 1976.
[19] T. Gilb, D. Graham, Software Inspections, Addison-Wesley, ISBN 0-201-63181-4, 1993.
[20] W. S. Humphrey, Managing the Software Process, Reading, Mass., Addison-Wesley Publishing Co., 1989.
[21] W. S. Humphrey, A Discipline for Software Engineering, Reading, Mass., Addison-Wesley Publishing, 1995.
[22] IEEE Standards Collection, Software Engineering, Std 830-1993, 1994.
[23] C. M. Judd, E. R. Smith, L. H. Kidder, Research Methods in Social Relations, Hartcourt Brace Jovanovich Inc., 6th edition, 1986.
[24] E. Kamsties, C. M. Lott, An Emperical Evaluation of Three Defect Detection Techniques, Proceedings of the Fifth European Software Engineering Conference, September 1995.
[25] O. Laitenberger, J.M. DeBaud, Perspective-based Reading of Code Documents at Robert Bosch GmbH, Conference Proceedings: Empirical Assessment and Validation in Software Engineering, Keele, UK., March 1997.
[26] L. P. W. Lau, C. Sauer, R. Jeffrey, Validating the Defect Detection Performance Advantage of Group Designs for Software Reviews: Report of a Laboratory Experiment Using Program Code, Centre for Advanced Empirical Studies (CAESAR) Technical Report 96/8, Sydney, 1996.
[27] R. C. Linger, H. D. Mills, B. Witt, Structured Programming: Theory and Practice, in The Systems Programming Series, Addison Wesley, 1979.
[28] C. M. Lott, H. D. Rombach, Repeatable Software Engineering Experiments for Comparing Defect-Detection Techniques, Empirical Software Engineering: An International Journal, vol. 1, no. 2, 241-277, 1996.
[29] Y. Mashiko and V. R. Basili, Using the GQM paradigm to Investigate Influential Factors for Software Process Improvement, Journal of Systems and Software, Vol. 36, pages 17-32, 1997.
[30] P. McCarthy, A. Porter, H. Sih, L. Votta, An Experiment to Assess Cost Benefits of Inspection Meetings and their Alternatives: A Pilot Study, Proceedings of the Third International Metrics Symposium, Berlin, Germany, 1996.
[31] R. McGill, J. W. Tukey, W. Larsen, Variations of boxplots, The American Statistician, 32(1), 12-16.
[32] A. Porter, L. G. Votta, V. R. Basili, Comparing Detection Methods for Software Requirements Inspections: A Replicated Experiment, IEEE Transactions on Software Engineering, Vol. 21, No. 6, Jund pp. 563-575, 1995.
[33] C. Sauer, R. Jeffery, L. Lau, P. Yetton, A Behaviourally Motivated Programme for Empirical Research into Software Development Technical Reviews, Centre for Advanced Empirical Studies (CAESAR) Technical Report 96/5, 1996.
[34] G. Schneider, J. M. Michael, W. Tsai, An Experimental Study of Fault Detection in User Requirements Documents, ACM Transactions on Software Engineering and Methodology, vol. 1, no. 2, pp. 188-204, Apr. 1992.
[35] S. S. Shapiro, M. B. Wilk, An Analysis of Variance Test for Normality (concrete samples), Biometrika, 52: 591-611, 1965.
[36] S. Siegel, N. J. Castellan, Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill, 1988.
[37] L. G. Votta, Does every inspection needs a meeting?, Proceedings of ACM SIGSOFT '93 Symposium on Foundations of Software Engineering, Association of Computing Machinery, 1993.

Appendix A: Example of a Scenario

Scenario of a Tester

Perspective-based Reading

Perspective based reading is the concept that the various customers of a product should read a document in such a way as to find out if the document satisfies their needs for it. In doing so it is hoped that the reader will find defects and be able to asses the document from their particular point of view.

Test-based Reading

For each requirement or functional specification (item), make up a test or set of tests that will allow you to ensure that the implementation satisfies the requirement. Use your standard test approach and test criteria to make up the test suite. For each requirement or functional specification, ask yourself the following questions:

1. Do you have all the information necessary to identify the item being tested and to identify your test criteria? Can you make up reasonable test cases for each item based upon the criteria?
2. Is there another requirement or functional specification for which you would generate a similar test case but would get a contradictory result?
3. Can you be sure that the test you generated should yield the correct value in the correct units?
4. Are there other interpretations of this requirement that the implementor might make based upon the way the requirement or functional specification is defined? Will this effect the tests you make up?
5. Does the requirement or functional specification make sense from what you know about the application or from what is specified in the general description?

Appendix B: Results of the Analysis for defect Classes

WS 95/96 run

	Class	Observed	Expected	Total	Group	Class	p-value
hard defects	A	1	0.947	38	6	6	1.000
	E	1	0.316	38	6	2	0.291
	I	2	1.894	38	6	12	1.000
	O	2	2.842	38	6	18	0.756
	M	0	0.000	38	6	0	1.000
moderate	A	2	1.737	38	11	6	1.000
	E	1	0.579	38	11	2	1.000
	I	3	3.474	38	11	12	1.000
	O	5	5.211	38	11	18	1.000
	M	0	0.000	38	11	0	1.000
easy defects	A	3	3.316	38	21	6	1.000
	E	0	1.105	38	21	2	0.200
	I	8	6.632	38	21	12	0.566
	O	11	9.947	38	21	18	0.646
	M	0	0.000	38	21	0	1.000

Table 22. Analysis for the Designer Perspective

	Class	Observed	Expected	Total	Group	Class	p-value
hard defects	A	0	0.364	33	3	4	1.000
	E	0	0.091	33	3	1	1.000
	I	1	1.000	33	3	11	1.000
	O	2	1.455	33	3	16	0.652
	M	0	0.091	33	3	1	1.000
moderate	A	1	1.091	33	9	4	1.000
	E	1	0.273	33	9	1	0.273
	I	3	3.000	33	9	11	1.000
	O	3	4.360	33	9	16	0.581
	M	1	0.273	33	9	1	0.273
easy defects	A	3	2.545	33	21	4	1.000
	E	0	0.636	33	21	1	0.364
	I	7	7.000	33	21	11	1.000
	O	11	10.182	33	21	16	0.798
	M	0	0.636	33	21	1	0.364

Table 23. Analysis for the User Perspective

Perspective	Detection Rate	Found Defects (Total)	Class	Defect Detection Rate (Class)	Found Defects (Mean/ Class)	Expected	p-value
Designer	0.3061	17.14	A	0.2222	2.00	2.755	0.46978
			E	0.1111	0.33	0.918	0.55784
			I	0.3889	6.22	4.898	0.41962
			O	0.3362	8.40	7.653	0.66452
			M	0.0000	0.00	0.918	0.558
Tester	0.2361	13.22	A	0.2500	2.25	2.125	0.69539
			E	0.2778	0.83	0.708	1.00000
			I	0.2222	3.56	3.778	1.00000
			O	0.2619	6.55	5.903	0.81614
			M	0.2778	0.83	0.708	1.00000
User	0.2296	12.86	A	0.1313	1.18	2.066	0.69432
			E	0.1250	0.38	0.689	0.59119
			I	0.3281	5.25	3.674	0.39253
			O	0.2370	5.93	5.740	1.00000
			M	0.0625	0.19	0.689	0.591

Table 24. Binomial Test for Defect Detection on Defect Classes

WS 96/97 run

	Class	Observed	Expected	Total	Group	Class	p-value
hard defects	A	1	0.903	31	7	4	0.581
	E	0	0.000	31	7	0	1.000
	I	1	2.258	31	7	10	0.474
	O	4	3.613	31	7	16	1.000
	M	1	0.226	31	7	1	0.226
moderate	A	1	1.161	31	9	4	1.000
	E	0	0.000	31	9	0	1.000
	I	1	2.903	31	9	10	0.299
	O	7	4.645	31	9	16	0.268
	M	0	0.290	31	9	1	1.000
easy defects	A	2	1.935	31	15	4	1.000
	E	0	0.000	31	15	0	1.000
	I	8	4.839	31	15	10	* 0.058
	O	5	7.742	31	15	16	0.214
	M	0	0.484	31	15	1	1.000

Table 25. Analysis for the Designer Perspective

	Class	Observed	Expected	Total	Group	Class	p-value
hard defects	A	2	1.531	32	7	7	1.000
	E	1	0.438	32	7	2	0.390
	I	1	2.188	32	7	10	0.478
	O	2	2.625	32	7	12	0.750
	M	1	0.219	32	7	1	0.219
moderate	A	3	2.188	32	10	7	0.685
	E	1	0.625	32	10	2	1.000
	I	1	3.125	32	10	10	0.188
	O	5	3.750	32	10	12	0.534
	M	0	0.313	32	10	1	1.000
easy defects	A	2	3.281	32	15	7	0.460
	E	0	0.938	32	15	2	0.502
	I	8	4.688	32	15	10	* 0.054
	O	5	5.625	32	15	12	0.780
	M	0	0.469	32	15	1	1.000

Table 26. Analysis for the Tester Perspective

	Class	Observed	Expected	Total	Group	Class	p-value
hard defects	A	0	0.750	32	4	6	0.615
	E	0	0.125	32	4	1	1.000
	I	3	1.625	32	4	13	0.392
	O	1	1.500	32	4	12	1.000
	M	0	0.000	32	4	0	1.000
moderate	A	4	2.438	32	13	6	0.232
	E	1	0.406	32	13	1	0.406
	I	2	5.281	32	13	13	* 0.089
	O	6	4.875	32	13	12	0.564
	M	0	0.000	32	13	0	1.000
easy defects	A	2	2.813	32	15	6	0.691
	E	0	0.469	32	15	1	1.000
	I	8	6.094	32	15	13	0.406
	O	5	5.625	32	15	12	0.780
	M	0	0.000	32	15	0	1.000

Table 27. Analysis for the User Perspective

Perspective	Detection Rate	Found Defects (Total)	Class	Defect Detection Rate (Class)	Found Defects (Mean/ Class)	Expected	p-value
Designer	0.2096	11.74	A	0.1375	1.24	1.886	0.69520
			E	0.1667	0.00	0.629	0.60717
			I	0.2344	3.75	3.354	0.75769
			O	0.2589	6.47	5.240	0.63331
			M	0.0625	0.19	0.629	0.60717
Tester	0.2137	11.97	A	0.2778	2.50	1.923	0.69449
			E	0.1667	0.50	0.641	1.00000
			I	0.2500	4.00	3.419	0.75998
			O	0.1941	4.85	5.343	0.80675
			M	0.1111	0.33	0.641	1.00000
User	0.2489	13.94	A	0.1889	1.70	2.240	0.69923
			E	0.2222	0.67	0.747	1.00000
			I	0.3889	6.22	3.982	0.14262
			O	0.1847	4.62	6.223	0.48652
			M	0.0000	0.00	0.438	1.00000

Table 28. Binomial Test for Defect Detection on Defect Classes