The Empirical Investigation of Perspective-Based Reading

Victor R. Basili[1], Scott Green[2], Oliver Laitenberger[3],
Filippo Lanubile[1], Forrest Shull[1],
Sivert Sørumgård[5], Marvin V. Zelkowitz[1]

Keywords: perspective-based reading, reading technique, requirement specification, defect detection, experimental software engineering

Abstract

We consider reading techniques a fundamental means of achieving high quality software. Due to the lack of research in this area, we are experimenting with the application and comparison of various reading techniques. This paper deals with our experiences with a family of reading techniques known as Perspective-Based Reading (PBR), and its application to requirements documents. The goal of PBR is to provide operational scenarios where members of a review team read a document from a particular perspective, e.g., tester, developer, user. Our assumption is that the combination of different perspectives provides better coverage of the document, i.e., uncovers a wider range of defects, than the same number of readers using their usual technique.

To test the effectiveness of PBR, we conducted a controlled experiment with professional software developers from the National Aeronautics and Space Administration / Goddard Space Flight Center (NASA/GSFC) Software Engineering Laboratory (SEL). The subjects read two types of documents, one generic in nature and the other from the NASA domain, using two reading techniques, a PBR technique and their usual technique. The results from these experiments, as well as the experimental design, are presented and analyzed. Teams applying PBR are shown to achieve significantly better coverage of documents than teams that do not apply PBR.

We thoroughly discuss the threats to validity so that external replications can benefit from the lessons learned and improve the experimental design if the constraints are different from those posed by subjects borrowed from a development organization.

---

[1] University of Maryland, USA
[2] NASA Goddard Space Flight Center, USA
[3] University of Kaiserslautern, Germany
[4] University of Trondheim, Norway

1. Reading Scenarios

The primary goal of software development is to generate systems that satisfy the user's needs. However, the various documents associated with software development (e.g., requirements documents, code and test plans) often require continual review and modification throughout the development life cycle. In order to analyze these documents, reading is a key, if not *the* key technical activity for verifying and validating software work products. Methods such as inspections (Fagan, 1976) are considered most effective in removing defects during development. Inspections rely on effective reading techniques for success.

Reading can be performed on all documents associated with the software process and can be applied as soon as the documents are written. However, except for Mills' reading by step-wise abstraction (Linger, 1979), there has been very little written on reading techniques. Most efforts have been associated with methods that simply assume that the given document can be read effectively (e.g., inspections, walk-throughs, reviews), but techniques for reading particular documents, such as requirements documents or test plans, do not exist. In cases where techniques do exist, the required skills are neither taught nor practiced. In teaching program design, for example, almost all effort is spent learning how to *write* code rather than how to *read* code. Thus, when it comes to reading, little exists in the way of research or practice.

In the NASA/GSFC Software Engineering Laboratory (SEL) environment, we have learned much about the effectiveness of reading and reading-based approaches through the application and evaluation of methodologies such as Cleanroom. We are now part of a group (ISERN[6]) that has undertaken as one of its activities, a research program to define and evaluate software reading techniques to support the various review methods[7] for software development.

The work reported in this paper was conducted within the confines of SEL. The SEL, started in 1976, has been developing technology aimed at improving the process of developing flight dynamics software for NASA/GSFC. This software is typically written in FORTRAN, C, C++,

[6] ISERN is the International Software Engineering Research Network whose goal is to support experimental research and the replication of experiments.
[7] We define the terms "technique" and "method" as follows: A technique is a series of steps, producing some desired effect, and requiring skilled application. A method is a management procedure for applying software techniques, which describes not only how to apply a technique, but also under what conditions the technique is appropriate.

or Ada. Systems can range from 20K to 1M lines of source code, with development teams of up to 15 persons working over a one to two year period.

1.1 Scenario-Based Reading

Since we believe that software development and analysis techniques need to be context dependent, well-defined, goal-oriented, and demonstrated effective for purpose, we established the following goals for defining reading techniques:

- The technique should be associated with the particular document (e.g., requirements) and the notation in which the document is written (e.g., English text). That is, it should fit the appropriate development phase and notation.
- The technique should be tailorable, based upon the project and environment characteristics. If the problem domain changes, so should the reading technique.
- The technique should be detailed, in that it provides the reader with a well-defined process. We are interested in usable techniques that can be repeated by others.
- The technique should be specific in that each reader has a particular purpose or goal for reading the document and the procedures support that goal. This can vary from project to project.
- The technique should be focused in that a particular technique provides a particular coverage of the document, and a combination of techniques provides coverage of the entire document.
- The technique should be studied empirically to determine if and when it is most effective.

To this end, we have defined a set of techniques, which we call proactive process-driven scenarios, in the form of algorithms that readers can apply to traverse the document with a particular emphasis. Because the scenarios are focused, detailed, and specific to a particular emphasis or viewpoint, several scenarios must be combined to provide coverage of the document.

We have defined an approach to generating a family of reading techniques based upon operational scenarios, illustrated in Figure 1. An operational scenario requires the reader to first create a model of the product, and then answer questions based on analyzing the model with a particular emphasis. The choice of abstraction and the types of questions asked may depend on the document being read, the problem history of the organization or the goals of the organization.

So far, two different families of scenario-based reading techniques have been defined for requirements documents: perspective-based reading and defect-based reading. Defect-based reading was the subject of an earlier set of experiments. Defect-based reading was defined for reading documents written in SCR style (Heninger, 1980), a formal notation for event-driven process control systems, and focuses on different defect classes, e.g., missing functionality and data type inconsistencies. These create three different scenarios: data type consistency, safety properties, and ambiguity/missing information. An experimental study (Porter, 1995) analyzed defect-based reading, ad hoc reading and checklist-based reading to evaluate and compare them with respect to their effect on defect detection rates. Major results were that (1) scenario readers performed better than ad hoc and checklist readers with an improvement of about 35%, (2) scenarios helped reviewers focus on specific defect classes but were no less effective at detecting other defects, and that (3) checklist reading was no more effective than ad hoc reading. However, the experiment discussed in this paper is concerned with an experimental validation of perspective-based reading, and so we treat it in more detail in the next section.

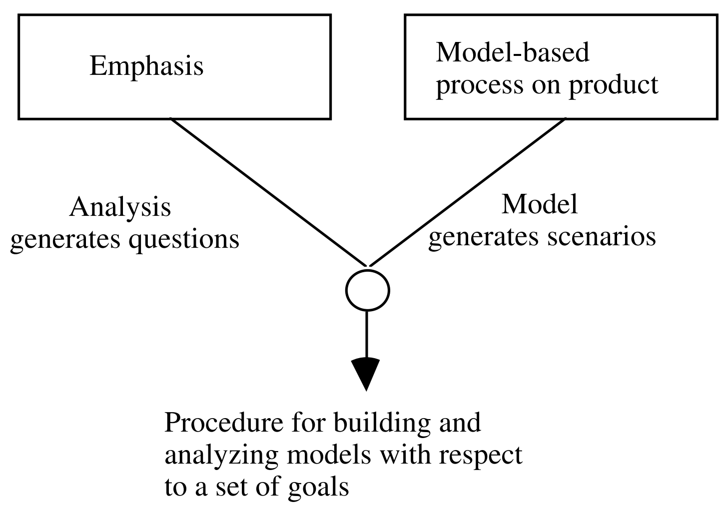

Figure 1. Building focused, tailored reading techniques.

1.2 Perspective-Based Reading

Perspective-based reading (PBR) focuses on the point of view or needs of the customers or consumers of a document. For example, one reader may read from the point of view of the tester, another from the point of view of the developer, and yet another from the point of view of the user of the system. To provide a proactive scenario, each of these readers produces some physical model which can be analyzed to answer questions based upon the perspective. For example, the team member reading from the perspective of the tester would design a set of tests for a potential test plan and answer questions arising from the activities being performed. Similarly, the team member reading from the perspective of the developer would generate a high level design, and the team member representing the user would create a user's manual. Each scenario is focused on one perspective. The assumption is that the union of the perspectives provides extensive coverage of the document, yet each reader is responsible for a narrowly focused view of the document, which should lead to more in-depth analysis of any potential errors in the document.

Consider, as an example, the procedure for a reader applying the test-based perspective to a requirements specification document:

> **Reading Procedure**: For each requirement, 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. While making up your test suite for each requirement, 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 for which you would generate a similar test case but would get a contradictory result?
> 3. Can you be sure the test you generated will 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 is defined? Will this effect the test you made up?
> 5. Does the requirement make sense from what you know about the application and from what is specified in the general description?

These five questions form the basis for the approach that the test-based reader will use to review the document.

We are proposing two different series of experiments for evaluating perspective-based reading techniques. The first series of experiments, the subject of this current paper, is aimed at discovering if perspective-based reading is more effective than current practices. We tested this using professionals within the SEL. It is expected that other studies will be run in different environments using the same artifacts where appropriate. A second series, to be undertaken later, will be used to discover under which circumstances each of the various perspectives is most effective.

1.3 Experimental Plan

Our method for evaluating PBR was to compare its effectiveness in uncovering defects with the approach people were already using for reading and reviewing requirements specifications. Thus, it assumes some experience in reading requirements documents on the part of the subjects. More specifically, the current NASA SEL reading technique (SEL, 1992) had evolved over time and was based upon recognizing certain types of concerns which were identified and accumulated as a set of issues requiring clarification by the document authors, typically the analysts and users of the system.

To test our hypotheses concerning PBR, a series of partial factorial experiments was designed, in which subjects would be given one document and told to discover defects using their current method. They would then be trained in PBR and given another document in order to see if their performance improved. The main research question was:

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

Our hypothesis is that "the union of the defects detected by groups of individuals with unique PBR roles provides a greater coverage of the documents than the union of defects detected by groups using the usual NASA technique."

As by-products of the main research question, we were also interested in the following secondary questions:

- If individuals read a document using PBR, would a different number of defects be found than if they read the document using their usual technique?
- Does a reviewer's experience in the role (designer, tester, user) influence performance when using PBR?

While we were interested in the effectiveness of PBR within our SEL environment, we were also interested in the general applicability of the techniques in environments different from the flight dynamics software that the SEL generally builds. Thus two classes of documents were developed: a domain-specific set that would have limited usefulness outside of NASA, and a generic set that is more representative of other domains and could be reused in other contexts.

For the NASA flight dynamics application domain, two small specifications derived from an existing set of requirements documentation were used. These specification documents, seeded with errors common to the environment, were labeled NASA_A and NASA_B. For the generic application domain, two requirements documents were developed and seeded with errors. These applications were an automated parking garage (PG) control system, and an automated bank teller machine (ATM).

1.4. Structure of this Paper

In section 2, we give a short discussion of the experimental design that we employed to test the effectiveness of PBR in the SEL environment and give a short overview of how we conducted two runs of this experiment. Section 3 presents the analysis of the data we obtained in the experiment. Section 4 discusses the various threats to the validity of our results. We describe those threats that we were able to anticipate in the experimental design and address in our results. We also discuss several threats that we were unable to foresee, and the impact of those new threats on our results. Section 5 discusses our experiences regarding designing and carrying out the experiment. Finally, Section 6 summarizes our findings and concludes with some indications of future directions for this research.

2. Design of the Experiment

Two runs of the experiment were conducted. Due to our experiences from the initial run, some modifications were introduced in the second run. We therefore view the initial run as a pilot study to test our experimental design, and we have run the experiment once more under more appropriate testing conditions.

For both runs, the population was software developers from the NASA SEL environment. All subjects were volunteers so we did not have a random sample population. We accepted everyone who volunteered, and nobody participated in both runs of the experiment.

2.1 Factors in the Design

In designing the experiment, we had to consider what factors were likely to have an impact on the results. The experimental design takes these *independent variables* into account and allows each of them to be separable from the others in order to allow for testing a causal relationship to the defect detection rate, the *dependent variable* under study.

Below we list the independent variables that we could manipulate in each separate run of the experiment.

- **Reading technique:** We have two alternatives: One is using a PBR technique, and the other is using the technique currently used for requirements document review in the NASA SEL environment, which we refer to as the “usual” technique.
- **Perspective:** Within PBR, a subject uses a technique based on one of the review perspectives. For this experiment we used the three perspectives previously described: Designer, Tester and User.
- **Requirements documents:** For each task to be carried out by the subjects, a requirements specification is handed out to be read and reviewed. The document will presumably have an impact on the results due to differences in size, domain and complexity.

There will also be other factors present that may have an impact on the outcome of the experiment, but that are hard to measure and control. These will be discussed in Section 4.

2.2 Constraints and Limitations

In designing the experiment we also took into account various constraints that restrict the way we could manipulate the independent variables. There are basically two factors that constrain the design of this experiment: time and cost.

- **Time:** Since the subjects in this experiment are borrowed from a development organization, we could not expect to have them available for an indefinite amount of time. This required us to make the experiment as time-efficient as possible without compromising the integrity of the design.
- **Cost:** For the same reason, we could not get as many subjects as we would have liked. Given the salaries typical for experienced software designers, we estimated the burdened cost (i.e., with overhead) per individual would be at least $500 per day, or over $1000 per participant.

This did not include the costs of the experimenters to set up, run, and analyze the results. For this reason, a major constraint in the experimental design would be to achieve meaningful results with a minimal number of subjects and a minimal number of replications. Specifically, we knew that we could expect to get between 12 and 18 subjects for two days on any run of the experiment.

Since we had to rely on volunteers, we had to provide some potential benefit to the subjects and the organization that was supporting their participation. Training in a new approach provided some benefit for their time. This had an impact on our experimental design because we had to treat people equally as far as the training they received.

2.3 Choosing a Design

Due to the constraints, we found constructing real teams of three reviewers to work together in the experiment to be unfeasible. With twelve subjects, we would only have four teams that would allow for only two treatments (use of PBR and the usual technique) with two data points in each. In order to achieve more statistical validity, we had each reviewer work independently, yielding six data points for each treatment. This decision not to use teams was supported by similar experiments (Parnas, 1985) (Porter, 1995) (Votta, 1993), where the team meetings were reported to have little effect in terms of the defect coverage; the meeting gain was outweighed by the meeting loss. Therefore we present no conclusions about the effects of PBR team meetings in practice. We do however examine the defect coverage that can result from teams by grouping reviewers into simulated teams which combine one reviewer from each of the perspectives. We discuss this further in Section 3.1.

The tasks performed by the subjects consisted of reading and reviewing a requirements specification document and recording the identified defects on a form. The treatments, which had the purpose of manipulating one or more of the independent variables, were aimed at teaching the subjects how to use PBR. There were four ways that we could have arranged the order of tasks and treatments for a group of subjects:

1. Start by teaching PBR, then do all tasks using PBR (experimental group)
2. Do all tasks using the usual technique (control group)
3. Start by teaching PBR, then do some tasks using PBR, followed by tasks using the usual technique.
4. Do pre-task(s) with the usual technique, then teach PBR, followed by post-task(s) using PBR.

In the first two options, the reading technique is a between-groups factor, in which each subject participates in only one treatment condition, either PBR or the usual technique. These two options were rejected because of the limited number of participants in the experiment. Furthermore, the control group of volunteers would not benefit from this study since they would not be learning anything about PBR. We did not believe this was appropriate given the support we got from the development organization.

In the other two options, the reading technique is a repeated-measures factor, in which each subject provides data under each of the treatment conditions. Each subject serves as its own control because each subject uses both PBR and the usual technique. These two last options are more efficient than the first two because they double the number of available observations. The experiment is also more attractive in terms of getting subjects, since they would all receive similar training.

Option 3, where the subjects first use PBR and then switch to their usual technique, was not considered a viable alternative because their recent knowledge in PBR may have undesirable influences on the way they apply their usual technique. The opposite may also be true, that their usual technique has an influence on the way they apply PBR. However, a prescriptive technique, such as a scenario-based reading technique, is assumed to produce a greater carry-over effect than a non-prescriptive technique, such as the usual technique at NASA. An analogous decision was taken in a related experiment (Porter, 1995) where subjects starting with a defect-based reading technique continued to apply it in the second experimental task. Thus, option 4 was selected.

All documents reviewed by a subject must be different. If a document was reviewed more than once by the same subject, the results would be disturbed by the subject's non-erasable knowledge about defects found in previous readings. This meant that we had to separate the subjects into two groups - one reading the first document and one reading the second in order to be able to compare a PBR and a usual reading of a document.

Based on the constraints of the experiment, each subject would have time to read and review no more than four documents: two from the generic domain, and two from the NASA domain. In addition, we needed one sample document from each domain for training purposes. We ended up providing the following documents:

- **Generic:**
  - Automatic teller machine (ATM) - 17 pages, 29 seeded defects.
  - Parking garage control system (PG) - 16 pages, 27 seeded defects.
  - Video rental system - 14 pages, 16 seeded defects (for training)
- **NASA:**
  - Flight dynamics (NASA_A) - 27 pages, 15 seeded defects
  - Flight dynamics (NASA_B) - 27 pages, 15 seeded defects
  - NASA sample - 9 pages, 6 seeded defects (for training)

Since we have sets of different documents and techniques to compare, it became clear that a variant of factorial design would be appropriate. Such a design would allow us to test the effects of applying both of the techniques on both of the relevant documents. A full factorial design would be inappropriate for two reasons: (1) It would require some subjects to apply the ordering of techniques that we previously argued against, and (2) It would require each subject to use all three perspectives at some point. Given our constraints, this would require an excessive amount of training, and perhaps even more important, the perspectives would likely interfere with each other, causing an undesirable learning effect.

We blocked the design on technique, perspective, document and reading sequence in order to get an equal distribution of the values of the different independent variables. Thus we ended up with two groups of subjects, where each group contains three subgroups, one for each perspective (see Figure 2). Our goal was to have a minimum of 12 subjects, yielding six for group 1 (with two in each perspective) and six for group 2 (again with two in each perspective).

<table>
<tr><td></td><td colspan="3">Group 1</td><td colspan="3">Group 2</td><td></td></tr>
<tr><td></td><td>Designer</td><td>Tester</td><td>User</td><td>Designer</td><td>Tester</td><td>User</td><td></td></tr>
<tr><td rowspan="4">usual technique</td><td colspan="3">Training</td><td colspan="3">Training</td><td rowspan="4">First day</td></tr>
<tr><td colspan="3">NASA A</td><td colspan="3">NASA B</td></tr>
<tr><td colspan="3">Training</td><td colspan="3">Training</td></tr>
<tr><td colspan="3">ATM</td><td colspan="3">PG</td></tr>
<tr><td rowspan="5">PBR technique</td><td colspan="6">Teaching of PBR</td><td rowspan="5">Second day</td></tr>
<tr><td colspan="3">Training</td><td colspan="3">Training</td></tr>
<tr><td colspan="3">PG</td><td colspan="3">ATM</td></tr>
<tr><td colspan="3">Training</td><td colspan="3">Training</td></tr>
<tr><td colspan="3">NASA B</td><td colspan="3">NASA A</td></tr>
</table>

Figure 2. Design of the experiment.

2.4 Conducting the Experiment

We conducted the first run in November, 1994, with 12 subjects. Each run of the experiment took two days, a Monday where each subject used the usual technique to review a NASA and then a generic document, and a Wednesday where each subject was taught one of the three PBR perspectives and applied that to two additional documents.

After analyzing the results (discussed in the next section), we held a meeting with the subjects to give them our conclusions and to obtain feedback from them on how the experiment was conducted. Several potential problems were reported:

1. We tried to assign subjects to each perspective according to their experiences. However, we had mostly software designers and did not have an equitable breakdown of testers and users for our three perspectives. We decided that we would randomize assignments in any future run of the experiment.
2. The NASA documents (at 27 pages) were deemed too long for appropriate analysis in a single session. We decided to revise these into shorter documents for any future run of the experiment.
3. We gave each subject up to three hours to review each document (i.e., one document in the morning, and one in the afternoon). Only one subject took more than two hours, so it was agreed that a two hour limit per document would be as effective in any future replications.
4. The initial run included training sessions only for the generic documents, but the subjects felt training for the NASA documents was warranted as well. Therefore in subsequent runs, we needed training sessions before each document review. For this purpose we generated an additional sample document representative of the NASA domain.
5. Subjects were allowed to work at their own desks, subject to interruptions, telephone calls, and other hazards of office life as long as they kept a log of time actually spent on the experiment. For any replication we believe that there would be greater internal validity in the results if we used a more uniform setting. While the first run took place at the facility of the developer, we decided to use a classroom setting at the University of Maryland for any subsequent runs.
6. We found a few inconsistencies in our specifications that we did not anticipate. A few sentences were changed to make them less ambiguous. Most importantly, the specification consisted of a general description of the problem and then a series of precise specifications, some of which were intentionally incorrect. We decided that for any replication the general

description should be absolutely correct in order for the subjects to have some basis for making decisions. This required minor changes to some of the specifications.

Because of these changes, we decided that it was best to call the November, 1994 run a pilot study for our experimental design. We conducted a second run of the experiment in June of 1995 with 14 subjects. Since one of the 14 was not familiar with NASA flight dynamics applications, we only used the 13 other subjects in analyzing NASA_A and NASA_B. In the next section we present an analysis of both the pilot study and the June, 1995 run.

After the 1995 run of the experiment, we marked all reviews with respect to their defect detection rate. Each review was graded by two individuals who did not know whether the subject was using PBR or the usual technique in order to eliminate any potential bias in the grading. After several iterations of discussion and re-marking, we arrived at a set of defect lists that were considered representative of the documents. Since these lists were slightly different from the lists that were used in the pilot study, we re-marked all the reviews from November 1994 in order to make all results consistent. Our initial measure of defect coverage was the percentage of the seeded defects that was found by each reviewer.

3. Statistical Analysis

After the pilot study and 1995 run, we have a substantial base of observations from which to draw conclusions about PBR. This task is complicated, however, by the various sources of extraneous variability in the data. Specifically, we identify four other variables (besides the reading technique) which may have an impact on the detection rate of a reviewer: the experiment run within which the reviewer participated, the problem domain, the document itself, and the reviewer's experience.

The experiment run is taken into account by performing a separate analysis for the pilot study and the 1995 run. The domain is taken into account in a similar way, by performing separate analyses for generic and NASA documents. The technique used and the document read are represented by nominal-scale variables used in our models. We measured reviewer experience as the number of years the reviewer had spent performing jobs related to the assigned perspective.

We are also careful to note that there are variables that our statistical analysis cannot measure. Perhaps most importantly, an influence due to a learning effect would be hidden within the effect of the reading technique. The full list of these threats to validity is found in Section 4, and any interpretation of results must take them into account.

In Section 3.1, we analyze the coverage that could result from PBR review teams, by simulating teams composed of one reviewer from each perspective. Section 3.2 presents the analysis of scores on an individual basis. Section 3.3 takes an initial look at the analysis with respect to the reviewer perspectives. In Section 3.4, we analyze the relationship between the reviewer's experience and the individual scores. In each section, we present the general analysis strategy and some details on the statistical tests, followed by the statistical results.

3.1 Analysis for Teams

In this section, we give a preliminary analysis concerning our primary hypothesis of the effect of PBR on inspection teams. It should be noted that since the PBR techniques are specific, any one of them will not cover the entire document. It is assumed that several of them need to be combined, in a team, to offer complete coverage of the document. In composing these teams, we might select one reader from each of the perspectives. Or, we could use the expected error profile of a project to help determine the number and types of perspectives to be used. For example, if we knew that, for a particular application domain, there is a tendency to commit a large number of errors of omission and that the user perspective offers the most opportunity for exposing omission errors, we might create a team with a larger number of use-based readers. One interesting direction for future research will be in developing ways to tailor the selection of reviewer perspectives to the problem at hand. However, for this experiment we examine the defect coverage of teams composed of one reviewer from each of the three perspectives.

Because we are concerned only with the range of a team's defect coverage, and not with issues of how team members will interact, we simulate team results by taking the union of the defects detected by the reviewers on the team. We emphasize that this grouping of individual reviewers into teams was performed after the experiment's conclusion, and does not signify that the team members actually worked together in any way. The only real constraint on the makeup of a team which applied PBR is that it contain one reviewer using each of the three perspectives; the non-PBR teams can have any three reviewers who applied their usual technique. At the same time, the way in which the teams are composed has a very strong effect on the team scores, so an arbitrary choice can have a significant effect on the test results.

For these reasons, we used a permutation test to test for differences in team scores between the PBR and the usual technique. We examine hypothetical teams from both the pilot study and the 1995 run, but keep the analysis of each run separate. Since the generic and NASA problem domains are also very different, we compare reviewer scores on documents within the same domain only. This gives us four iterations of the permutation test, for which we present an informal description here.

In Section 2.3 we explained how reviewers were categorized into two groups, depending on which technique they applied to which document. Reviewers in Group 1 applied their usual technique to Document A and PBR to Document B, where Document A and Document B represent the two documents within either of the domains. We can thus generate the set of all possible non-PBR teams for Document A and the set of all possible PBR teams for Document B, and examine the defect coverage that could be expected from each technique by taking the average detection rate of each set. This ensures that our results are independent of any arbitrary choice of team members, but because the data points for all possible teams are not independent (i.e., each reviewer appears multiple times in this list of all possible teams), we cannot run simple statistical tests on these average values. For now, let us call these averages A_{1USUAL} and B_{1PBR}. We can then perform the same calculations for Group 2, in which reviewers applied their usual technique to Document B and PBR to Document A, in order to obtain averages A_{2PBR} and B_{2USUAL}. The test statistic

$$(A_{2PBR} - A_{1USUAL}) + (B_{1PBR} - B_{2USUAL})$$

then gives us some measure of how all possible PBR teams would have performed relative to all possible usual technique teams, for each document. For each test performed, Figure 3 shows the average of the PBR scores (A_{2PBR} and B_{1PBR}) against the average usual technique scores (A_{1USUAL} and B_{2USUAL}).

Now suppose we switch a reviewer in Group 1 with someone from Group 2. The new reviewer in Group 1 will be part of a usual technique team for document A even though he used PBR on this document, and will be part of a PBR team for Document B even though he applied the usual technique. A similar but reversed situation awaits the reviewer who suddenly finds himself in Group 2. If the use of PBR does in fact improve team detection scores, one would intuitively expect that as the PBR teams are diluted with usual technique reviewers, their average score will decrease (that is, A_{2PBR} and B_{1PBR} decrease), even as the average score of usual technique teams with more and more PBR members is being raised (that is, A_{1USUAL} and B_{2USUAL} increase). Thus, the test statistic computed above will decrease. On the other hand, if PBR does in fact have no effect, then as reviewers are switched between groups the only effect will be due to random effects, and team scores may improve or decrease with no correlation with the reading technique of the reviewers from which they are formed. So, let us now compute the test statistic for all possible permutations of reviewers between Group 1 and Group 2, and rank each of these scenarios in decreasing order by the statistic. We can now formulate the null hypothesis we are testing as:

> **H_0:** There is no difference in the defect detection rates of teams applying PBR as compared to teams applying the usual technique. That is, every successive dilution of a PBR team with non-PBR reviewers has only random effects on team scores.

The alternative hypothesis is:

> **H_a:** The defect detection rates of teams applying PBR are higher compared to teams using the usual technique. That is, every time the PBR teams were diluted with non-PBR reviewers they tended to perform somewhat worse relative to the usual technique teams.

If the scenario in which no dilution has occurred appears toward the top of the list (in the top 5%) we will reject H_0 and conclude that PBR does have a beneficial effect on team scores. Note that this is meant to be only a very rough and informal description of the intuition behind the test; the interested reader is referred to Edington's *Randomization Tests* (Edington, 1987).

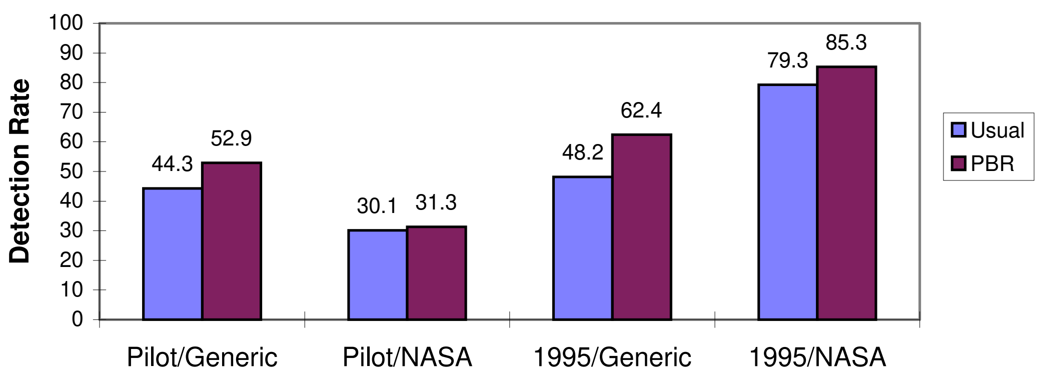

Figure 3. Simulated scores for undiluted teams

Table 1 summarizes the results. The first row of the table, for example, shows that twelve reviewers read the generic documents in the pilot study; there are 924 distinct ways they can be assigned into groups of 6. The group in which there was no dilution had the 61st highest test statistic, corresponding to a p-value of 0.0660. Both domains in the 1995 run had results significant at the 0.05 level, that is, we can reject H_0 because undiluted teams appear in the top 5% of all possible permutations between groups. The generic domain in the pilot study had results significant at the 0.1 level. However, since there were only 924 permutations generated in the pilot study, the power of the test is correspondingly less and it may be reasonable to reject the null hypothesis at the 0.1 level.

Experiment Run/ Domain	**Number of Group Permutations Generated**	**Rank of Undiluted Group**	**P-value**
Pilot/Generic	924	61	0.0660
Pilot/NASA	924	401	0.4340
1995/Generic	3003	2	0.0007
1995/NASA	1716	67	0.0390

Table 1. Results of permutation tests for team scores.

3.2 Analysis for Individuals

Although our main hypothesis was an investigation of teams, we decided to also analyze the results on an individual basis. The purpose was to determine whether an individual performed differently when reviewing with PBR than when using the usual technique. The dependent variable was again the defect rate, in this case, the percentage of true defects found by a single reviewer with respect to the total number of defects in the inspected document.

As we did for the analysis of team scores, we analyzed separately both the experiment runs (pilot study and 1995 run), and the problem domains (generic and NASA). Thus, we performed four separate analyses for each combination of the experiment run and problem domain.

For each analysis, the corresponding design is a 2×2 factorial experiment with repeated measures in blocks of size 2 (Winer, 1991). This analysis involves two factors, or treatments, on which there are repeated measures: the reading technique (*RTECH*) and the document reviewed (*DOC*). Both the dependent variables are on a nominal scale. The reading technique has two levels: PBR and usual. Also the document reviewed can assume two levels which depend on the problem domain considered: if generic domain, ATM and PG; if NASA domain, NASA_A and NASA_B.

Groups of subjects were assigned in two groups, or blocks, and had repeated measures on the two treatments within each block. The result is shown in Figure 4. Group 1 applied the usual technique to the ATM document and PBR to the PG document, while Group 2 applied PBR to ATM and the usual technique to the PG document. On the other hand, for the NASA problem domain, subjects in Group 1 read NASA_A document with their usual technique and NASA_B document with PBR, while subjects in Group 2 read the documents in the opposite fashion.

Generic domain		**NASA domain**	
<u>Group 1</u>	<u>Group 2</u>	<u>Group 1</u>	<u>Group 2</u>
usual/ATM	usual/PG	usual/NASA_A	usual/NASA_B
PBR/PG	PBR/ATM	PBR/NASA_B	PBR/NASA_A

Figure 4. 2×2 factorial experiments with repeated measures in blocks of size 2

These plans use all the treatment conditions required for the complete factorial experiment, but the block size is reduced from four to two; that is, within any block only two treatment combinations appear instead of the four possible treatment combinations. The cost of this reduction in block size is the loss of some information on interactions. Precisely, the interaction $RTECH \times DOC$ is totally confounded with the group main effect. This means that we cannot estimate the two-factor interaction separately from the group effect. However, we do not expect this interaction to be important because both the documents are within the same problem domain. In exchange, both of the main effects are completely within-block effects, and thus independent from the subject variability.

The two design plans in the pilot study have 6 subjects in each group. On the contrary, the 1995 run shows an unbalanced situation. For the generic problem domain, there are 8 subjects in Group 1 and 6 subjects in Group 2, while for the NASA domain, there are 7 subjects in Group 1 and 6 subjects in Group 2. Thus, in order to perform the analysis of variance for unbalanced design, we used the GLM procedure in the SAS statistical package (SAS, 1989), which uses the method of least squares to fit general linear models.

The analysis of variance for the 2×2 factorial design with repeated measures in blocks of size 2 uses *F* ratios to test three hypotheses:

Group effect or *RTECH* $\times$ *DOC* interaction effect

> **H_0:** There is no difference between subjects in Group 1 and subjects in Group 2 with respect to their mean defect rate scores.
>
> **H_a:** There is a difference between subjects in Group 1 and subjects in Group 2 with respect to their mean defect rate scores.

Main effect *RTECH*

> **H_0:** There is no difference between subjects using PBR and subjects using their usual technique with respect to their mean defect rate scores.
>
> **H_a:** There is a difference between subjects using PBR and subjects using their usual technique with respect to their mean defect rate scores.

Main effect *DOC*

> **H_0:** There is no difference between subjects reading the ATM document (or NASA_A document) and subjects reading the PG document (or NASA_B document) with respect to their mean defect rate scores.
>
> **H_a:** There is a difference between subjects reading the ATM document (or NASA_A document) and subjects reading the PG document (or NASA_B document) with respect to their mean defect rate scores.

The analysis makes a number of assumptions, which we were careful to fulfill: The dependent variable is measured on a ratio scale, and the independent variables are nominal. Observations are independent. The values tested for each level of the independent variables are normally distributed; we confirmed this with the Shapiro-Wilk W Test (Shapiro, 1965). Also, the test assumes that the sources are homogeneous with regard to the two groups. However, we note that the test is robust against violations of this last assumption for data sets such as ours in which the number of subjects in the largest treatment group is no more than 1.5 times greater than the number of subjects in the smallest (Hatcher, 1994). The test also assumes that the sample must be obtained through random sampling; this is a threat to the validity of our experiment, as we must rely on volunteers for our subjects (see Section 4).

3.2.1 Analysis of Pilot Study in the Generic Problem Domain

The analysis, summarized in Table 2, failed to reveal a significant main effect both for the reading technique (p = 0.2148) and the document reviewed (p = 0.4068). The interaction between reading technique and document, which is totally confounded in the group effect, also proved to be non-significant (p = 0.5582).

The mean defect rates obtained for each level of reading technique and document are displayed in Figure 5. The defect detection rate for PBR reviewers is slightly higher (24.92) than for reviewers using their usual technique (20.58). Although this difference is not statistically significant, it represents a 21% improvement over the usual detection rate.

Source	**df**	**SS**	**MS**	**F**	**p>F**
Between subjects	11	1205.5			
Group or RTECH × DOC	1	42.67	42.67	0.37	0.5582
Error	10	1162.83	116.28		
Within subjects	12	803.01			
Reading Technique (RTECH)	1	112.67	112.67	1.75	0.2148
Document (DOC)	1	48.17	48.17	0.75	0.4068
Error	10	642.17	64.22		

Table 2. ANOVA summary table for pilot study in the generic problem domain

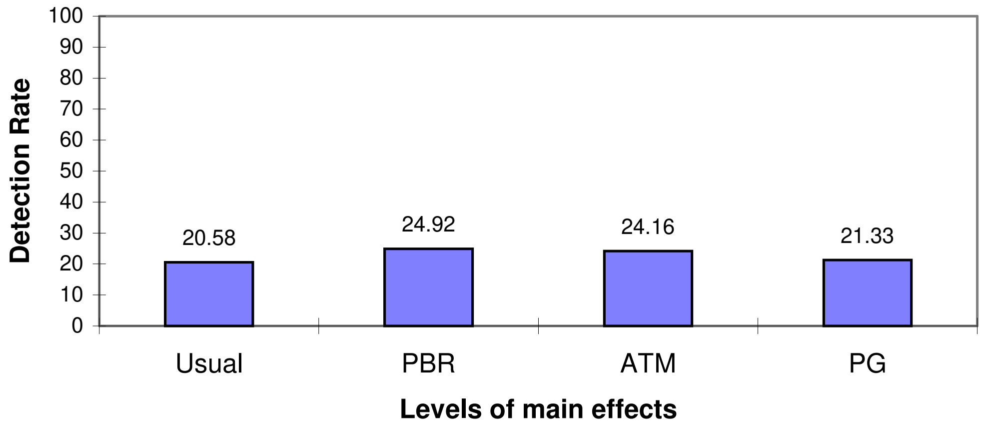

Figure 5. Individual mean scores of pilot study for the generic problem domain

3.2.2 Analysis of Pilot Study in the NASA Problem Domain

The analysis, summarized in Table 3, failed to reveal a significant main effect both for the reading technique ($p = 0.9629$) and the document reviewed ($p = 0.7109$). The interaction between reading technique and document, which is totally confounded in the group effect, also proved to be non-significant ($p = 0.1339$).

The mean defect rates obtained for each level of reading technique and document are displayed in Figure 6. Reviewers scored poorly regardless of which reading technique was used or which document was reviewed.

Source	df	SS	MS	F	p>F
Between subjects	11	1606.00			
Group or RTECH × DOC	1	337.50	337.50	2.66	0.1339
Error	10	1268.50	126.85		
Within subjects	12	744.01			
Reading Technique (RTECH)	1	0.17	0.17	0.00	0.9629
Document (DOC)	1	10.67	10.67	0.15	0.7109
Error	10	733.17	73.32		

Table 3. ANOVA summary table for pilot study in the NASA problem domain

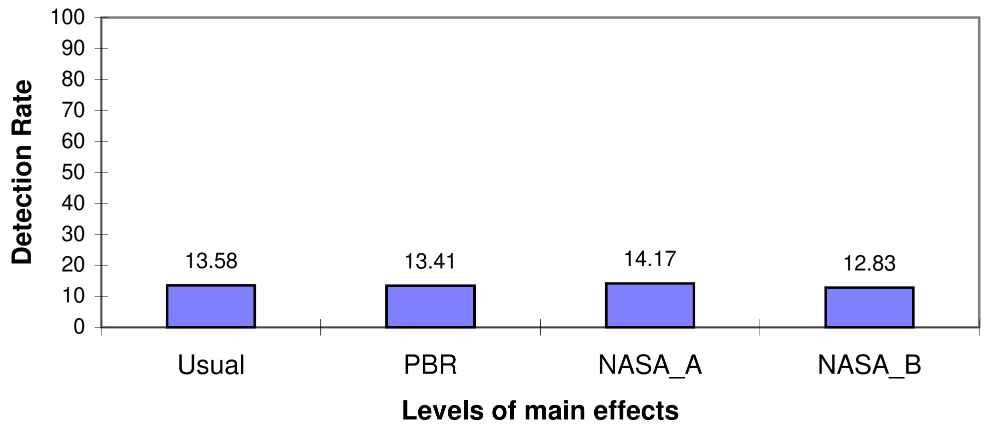

Figure 6. Individual mean scores of pilot study for the NASA problem domain

3.2.3 Analysis of 1995 Run in the Generic Problem Domain

The analysis, summarized in Table 4, revealed a significant main effect at the 0.05 level both for the reading technique ($p = 0.0019$) and the document reviewed ($p = 0.0160$). However, the defect detection rate has a stronger relationship with the reading technique ($R^2 = 0.44$) than with the document reviewed ($R^2 = 0.22$). R^2 indicates what percent of variance in the dependent variable is accounted for by the independent variable. Unlike the main effects, the interaction between reading technique and document, which is totally confounded in the group effect, proved to be non-significant ($p = 0.5213$).

The mean defect rates obtained for each level of reading technique and document are displayed in Figure 7. PBR reviewers (detection rate = 32.14) scored significantly higher than reviewers using their usual technique (detection rate = 24.64). Thus, there is 30% improvement over the usual detection rate. On the other hand, reviewers reading PG document (detection rate = 31.29) performed significantly better than reviewers reading ATM document (detection rate = 25.50).

Source	df	SS	MS	F	p>F
Between subjects	13	3093.17			
Group or RTECH × DOC	1	108.57	108.57	0.44	0.5213
Error	12	2984.60	248.72		
Within subjects	14	719.99			
Reading Technique (RTECH)	1	318.24	318.24	15.72	0.0019
Document (DOC)	1	158.81	158.81	7.84	0.0160
Error	12	242.94	20.24		

Table 4. ANOVA summary table for 1995 run in the generic problem domain

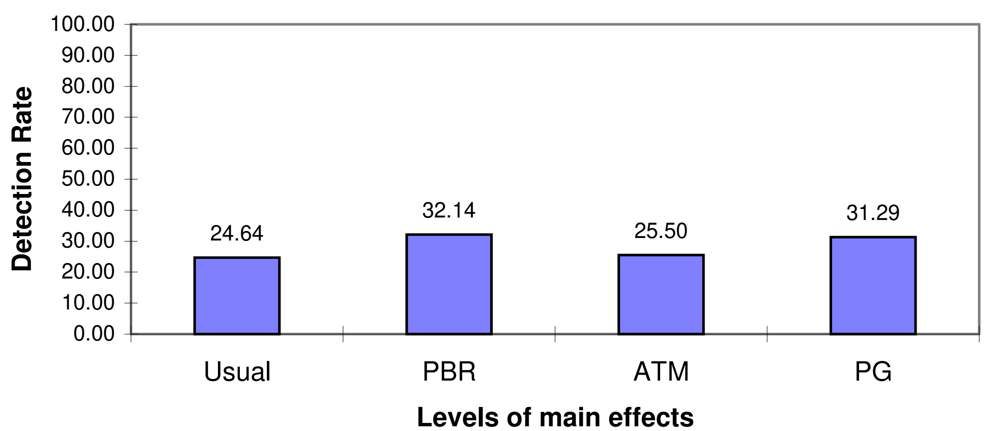

Figure 7. Individual mean scores of 1995 run for the generic problem domain

3.2.4 Analysis of 1995 Run in the NASA Problem Domain

The analysis, summarized in Table 5, failed to reveal a significant main effect both for the reading technique ($p = 0.4755$) and the document reviewed ($p = 0.9100$). The interaction between reading technique and document, which is totally confounded in the group effect, also proved to be non-significant ($p = 0.5394$).

Source	df	SS	MS	F	p>F
Between subjects	12	14405.61			
Group or RTECH × DOC	1	506.90	506.90	0.40	0.5394
Error	11	13898.71	1263.52		
Within subjects	13	2137.94			
Reading Technique (RTECH)	1	100.94	100.94	0.55	0.4755
Document (DOC)	1	2.48	2.48	0.01	0.9100
Error	11	2034.52	184.96		

Table 5. ANOVA summary table for 1995 run in the NASA problem domain

The mean defect rates obtained for each level of reading technique and document are displayed in Figure 8. The defect detection rate for PBR reviewers is slightly higher (51.23) than for reviewers using their usual technique (47.23). All the mean scores in the 1995 run were greatly better than the mean scores in the pilot study, thus confirming the need for improving the experimental conditions.

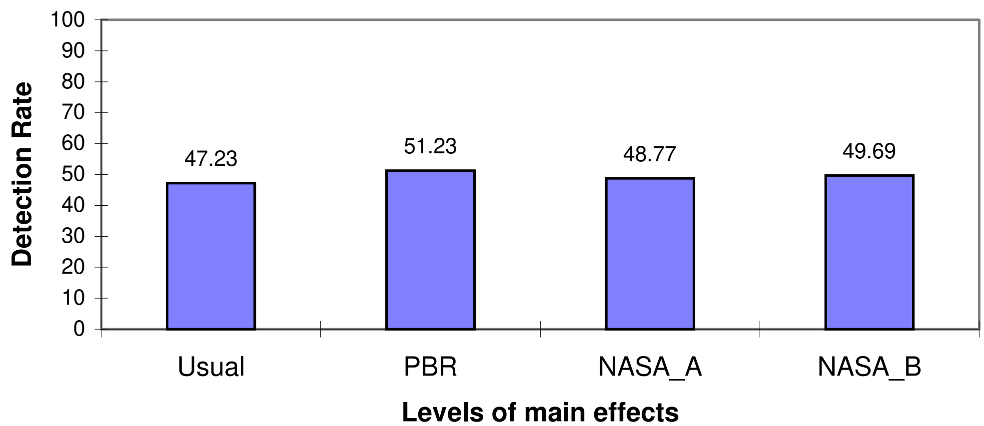

Figure 8. Individual mean scores of 1995 run for the NASA problem domain

3.3 Analysis for Perspectives

Aside from detection rates, we were also interested in determining if the PBR reviewers discovered a larger class of errors than those without PBR training and if the errors found were orthogonal (i.e., perspectives did not overlap in terms of the set of defects they helped detect). A full study of correlation between the different perspectives and the types and numbers of errors they uncovered will be the subject of future work, but for now we take a qualitative look at the results for each perspective by examining each perspective's coverage of defects and how perspectives overlap.

We formulate no explicit statistical tests concerning the detection rates of reviewers using each of the perspectives, but present Figures 9a and 9b as an illustration of the defect coverage of each perspective. Results within domains are rather similar; therefore we present the ATM coverage charts as an example from the generic domain and the document NASA_A charts as an example from the NASA domain. The numbers within each of the circle slices represent the number of defects found by each of the perspectives intersecting there. So, for example, ATM reviewers using the design perspective in the 1995 run found 11 defects in total: two were defects that no other perspective caught, three defects were also found by testers, one defect was also found by users, and five defects were found by at least one person from each of the three perspectives.

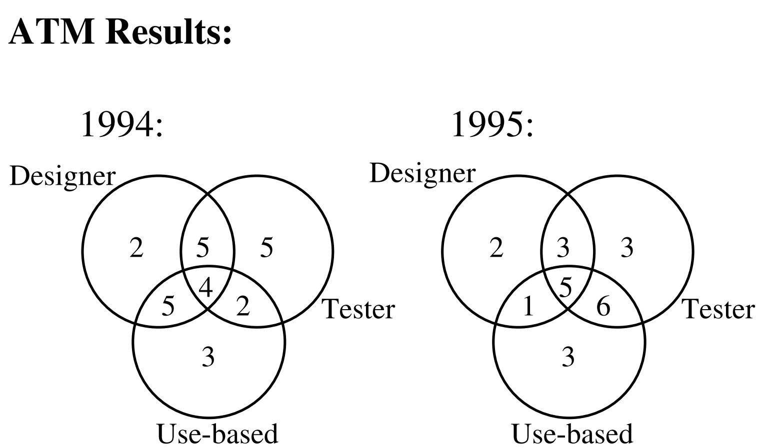

Figure 9a. Defect coverage for the ATM document in the generic domain

NASA_A Results:

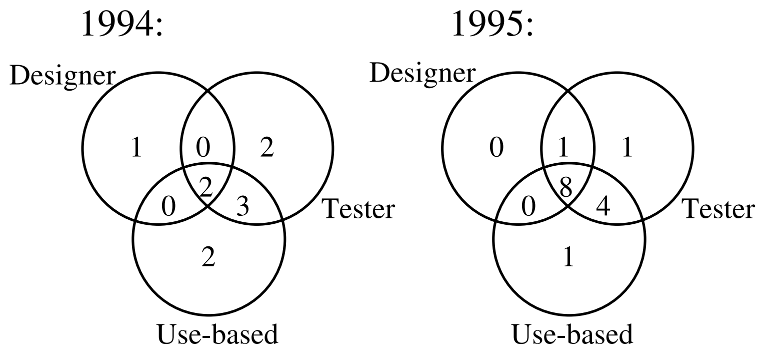

Figure 9b. Defect coverage for the NASA_A document in the NASA domain

3.4 Analysis for Reviewer's Experience

We measured reviewer experience via questionnaires used during the course of the experiment: a subjective question asked each reviewer to rate on an ordinal scale his or her level of comfort using such documents, and objective questions asked how many years the reviewer had spent in each of the perspective roles (designer, tester, user).

As shown in Figures 10a and 10b, the relationship between PBR defect rates and experience is weak. Reviewers with more experience do not perform better than reviewers with less experience. On the contrary, it appears that some less-experienced reviewers have learned to apply PBR better.

Both Spearman's and Pearson's correlation coefficients were computed in order to measure the degree of association between the two variables for each type of document in each experiment run, but in no case was there any value above 35% (values close to 100% would have indicated a high degree of correlation).

A similar lack of significant relationship has been found by Humphrey in analyzing the experiences with teaching the Personal Software Process (Humphrey, 1996).

Figure 10a. PBR defect rate versus role experience in the pilot study

Figure 10b. PBR defect rate versus role experience in the 1995 run

4. Threats to Validity

Threats to validity are factors beyond our control that can affect the dependent variables. Such threats can be considered unknown independent variables causing uncontrolled rival hypotheses to exist in addition to our research hypotheses. One crucial step in an experimental design is to minimize the impact of these threats. In this section, we present both those threats which we anticipated and tried to control for, as well those threats which were only realized after the fact.

We have two different classes of threats to validity: threats to *internal* validity and threats to *external* validity. Threats to internal validity constitute potential problems in the interpretation of the data from the experiment. If the experiment does not have a minimum internal validity, we can make no valid inference regarding the cause-effect relationship between independent and dependent variables. On the other hand, the level of external validity tells us nothing about whether the data is interpretable, but is an indicator of the generalizability of the results. Depending on the external validity of the experiment, the data can be assumed to be valid in other populations and settings.

4.1. Threats to Internal Validity

The following five threats to internal validity (Campbell, 1963) are discussed in order to reveal their potential interference with our experimental design:

- **History:** Since there was one day between the two days of the experiment, some of the improvement that appears due to technique may be attributed to other events that took place between the tests. The subjects were instructed not to discuss the experiment or otherwise do anything between the tests that could cause an unwanted effect on the results. We trusted the professional programmers in Group 1 not to discuss their specifications with the programmers in Group 2. Thus, we do not consider this effect to be very significant, but we cannot completely ignore it.
- **Maturation:** This is the effect of processes taking place within the subjects as a function of time, such as becoming tired or bored. But it may also be intellectual maturation, regardless of the experimental events. For our experiment, the likely effect would be that tests towards the end of the day tend to get worse results than they would normally. We provided long breaks between tests to try an avoid such a tendency. However, since the ordering of documents and domains was different for the two days, the differences between the two days may be disturbed by maturation effects. Looking at the design of the experiment, we see that an improvement from the first to the second day would be amplified for the generic documents, while it would be lessened for the NASA documents. Based on the results from the experiment, we see that this effect seems plausible. On the other hand, if we had chosen to have the same order of domains and documents in the two days, the threat would be worse because an improvement from the usual technique to PBR would be completely confounded with the maturation effect.
- **Testing:** Getting familiar with the tests may have effects on subsequent results. This threat has several components, including becoming familiar with the specifications, the technique, or the testing procedures. This effect may amplify the effects of the historical events and thus be part of the reason for improvement that has previously been considered a result of change

in technique. Testing effects may counteract maturation effects within each day. Although our subjects were already familiar with NASA documents, we tried to overcome unwanted effects by providing training sessions before each test where the subjects could familiarize themselves with the particular kind of document and technique. Also, the subjects received no feedback regarding their actual defect detection success during the experiment, so that it would presumably be difficult for them to discover whether aspects of their performance were in fact improving their detection rate or not. Furthermore, the generic documents are dissimilar enough that there little to be learned from the first document that could be transferred to the second. However, it would be interesting to replicate this experiment with a true control group who did not receive PBR training on day 2.

- **Instrumentation:** These effects are basically due to differences in the way of measuring scores. Our scores were measured by two people independently who did not know which treatment they were grading, and then discussed in order to resolve any disagreement consistently.
- **Selection:** For the 1995 run we used random assignment of subjects to perspective. Since PBR assumes the reviewers in a team use the perspectives with which they are familiar, the random assignment used in the experiment would presumably lead to an underestimation of the improvement caused by PBR.

Another threat to internal validity is the possibility that the subjects ignore PBR when they are supposed to use it. In particular, there is a danger that the subjects continue to use their usual technique. This need not be the result of a deliberate choice from the subject, but may simply reflect the fact that people unconsciously prefer to apply existing skills with which they are familiar. The only way of coping with this threat is to provide enhanced training sessions and some sort of control or measure of conformance to the assigned technique. However, as we have already shown, the PBR scenario did have a positive effect over the usual techniques in most of the documents, so even if the experiment was confounded with this effect, the true results would be stronger.

4.2. Threats to External Validity

Threats to external validity imply limitations to generalizing the results. The experiment was conducted with professional developers and with documents from an industrial context, so these factors should pose little threat to external validity. However, the limited number of data points is a potential problem which may only be overcome by further replications of the experiment. Other threats to external validity pertinent to the experimental design include (Campbell, 1963):

- **Interaction of testing and treatment:** A pretest may affect the subject's sensitivity of the experimental variable. Both of our groups receive similar pretests and treatments, so this effect may be of concern to us. We cannot avoid the fact that this is an experimental environment, and all subjects knew that. This, by itself, may affect the results and is a limitation of almost any experimental design.
- **Interaction of selection and treatment:** Selection biases may have different effects due to interaction with the treatment. One factor we need to be aware of is that all our subjects were volunteers. This may imply that they are more prone to improvement-oriented efforts than the average developer - or it may indicate that they consider the experiment an opportunity to get away from normal work activities for a couple of days. Thus, the effects can strike in either direction. Also, all subjects had received training in their usual technique, a property that developers from other organizations may not possess.
- **Reactive arrangements:** These effects are due to the experimental environment. In 1994, the pilot study was carried out in the subjects' own environment, and thus would be valid also in a real setting. We cannot assume the same for the 1995 results since this run was done in a classroom situation. However, the change of experimental environment between the experiment runs has made it easier to concentrate on the techniques and tests to be done, thus separating the techniques better.

Since this experiment was conducted using personnel from the NASA SEL environment, it is reasonable to discuss whether the results can be generalized to a NASA SEL context. This kind of generalization involves less of a change in context than is the case for an arbitrary organization; in particular the differences in populations can be ignored since the population for the experiment is in fact all of the NASA SEL developers.

Clearly, the results for the generic documents cannot be generalized to the NASA documents due to the difference in nature between the two sets of documents. The results for the NASA documents, on the other hand, may be valid since we used parts of *real* NASA documents.

5. Discussion

We have encountered problems in the two runs of the experiment which we have previously discussed. However, some of these problems are of a general nature and may be relevant in other experimental situations.

- *What is a good design for the experiment under investigation, given the constraints?*
  An important constraint in the design of our experiment was the necessity of a minimal number of subjects and experiment runs due to high costs and low availability of subjects. Unlike many other academic studies, it is simply not feasible to acquire another classroom full of students and rerun the experiment. This is not meant to disparage such studies (several of the authors have conducted numerous studies of that type in the past), but is only an indication that the more usual academic model is not applicable in the industrial setting we wish to study. Analytical techniques involving only a few subjects need to be developed and explored.

- *What is the optimal sample size? Small samples lead to problems in the statistical analysis while large samples represent major expenses for the organization providing the subjects.*
  Organizations generally have limits for the amount of subjects they are willing to part with for an experiment, so the cost concerns are handled by the organizations themselves. A small sample size requires us to be careful in the design in order to get as many useful data points as possible. In our case, we chose to neglect learning effects in order to avoid having control groups. While giving us more data points to be used in analyzing the difference between the two techniques, we remained uncertain as far as the threat to internal validity caused by learning effects is concerned.

- *We need to adjust to various constraints - how far can we go before the value of the experiment decreases to a level where it is not worthwhile?*
  The problem is how controlled can we make the experimental environment, yet still keep the industrial organization interested in participating. Giving each subject training in PBR and giving up a control group is an example of this tradeoff.

- *To what extent can experimental aspects such as design, instrumentation and environment be changed when the experiment still is to be considered a replication?*
  The software engineering literature contains many examples of experiments often under the guise of "case study." How often can we compare the results of two different case studies? In our own example, we viewed the 1994 pilot study and the 1995 run as distinct even though they were probably more similar than other published case studies. This is a non-trivial problem in how to build up a body of knowledge that others may reference.

- *What threats to validity did we fail to address?*

  The hardest part of any experiment is to admit what you did wrong and what did not go as planned. Throughout this report we addressed several aspects of our experimental design that could be improved. Some of the more important ones are:

  1. The maturation internal threat to validity may be a factor if tiredness in the afternoon affects results. Our experimental design favored the generic documents over the NASA documents, as we mentioned earlier.
  2. The testing internal threat to validity may be factor if there is a learning effect. Rerunning the experiment with a third control group, or rerunning the experiment to add a third or fourth day of PBR testing to see if there is further improvement after repeated PBR practice are ways to address this.
  3. Matching subject experience with PBR scenario would make the results more relevant to the usual NASA domain. Of course, rerunning this experiment within another domain would be necessary to generalize this technique outside of the NASA SEL flight dynamics domain.
  4. There are other issues besides coverage that are important when studying review teams. Even though we have indications that teams composed of developers using unique PBR techniques are better than teams using no formal reading technique, it is important to confirm these simulated results with a study in which real review teams are used.

A more fundamental problem that should be considered is to what extent the proposed technique actually is followed. This problem with process conformance is relevant in experiments, but also in software development where deviations from the process to be followed may lead to wrong interpretation of measures obtained. For experiments, one problem is that the mere action of controlling or measuring conformance may have an impact on how well the techniques work, thus decreasing the external validity.

Conformance is relevant in this experiment because there seems to be a difference that corresponds to experience level. Subjects with less experience seem to follow PBR more closely ("It really helps to have a perspective because it focuses my questions. I get confused trying to wear all the hats!"), while people with more experience were more likely to fall back to their usual technique ("I reverted to what I normally do.").

6. Conclusions and Future Directions

To get high quality software, the various documents associated with software development must be verified and validated. People doing the verification or validation effectively must get an understanding of the document. We consider reading the key technical activity to understand a document. In this paper we have presented a reading technique called Perspective-Based Reading (PBR) and its application to requirements documents.

We tested the effectiveness of PBR in two runs of a controlled experiment with professionals from the NASA SEL environment. The subjects used both their usual technique and PBR on generic requirements documents and on requirements documents from their application domain (NASA documents). As PBR can be used in methods, like inspections, where two or more reviewers of a review team individually look for defects, we were especially interested in the team results. Because of time and cost constraints, we simulated team results by combining the defects detected by the reviewers on the team.

In the first run (pilot study) of the experiment we only got significant results for teams using PBR on the generic documents. After the pilot study we made some changes to improve the experiment. In the 1995 run, PBR teams provided significantly better coverage of both NASA and generic documents. The reasons for this observed improvement, as compared to the pilot study, may include shorter assignments in the NASA problem domain and training session before each document review.

We have also compared PBR and the usual technique with respect to the individual performance of reviewers. Although in most cases, reviewers using PBR found about the same number of defects as reviewers using their usual technique, PBR reviewers did perform significantly better on the generic documents in the 1995 run. This is an unexpected result, since the true benefit of PBR is expected to be seen at the level of teams which combine several different perspectives for improved coverage. The results for individuals show that under certain conditions, the use of focused techniques may lead to improvements at the individual level as well

We think that better results could be achieved by more closely tailoring PBR to the specific characteristics of the NASA documents and NASA SEL environment. Partly, this conclusion is motivated by examining the distribution of discovered defects among the different PBR techniques: in generic documents there were a number of defects which were found by only one of the perspectives, while on NASA documents a much greater degree of overlap among the perspectives was observed. We also got feedback from the subjects that supported this view; several found it tempting to fall back to their usual technique when reading the NASA documents, thus underestimating the effect of using PBR. This observation is also supported by the lack of relationship between defect coverage using PBR and reviewer's experience in the assigned role.

A possible direction for further experimentation would be to do a case-study of a NASA SEL project to obtain more qualitative data, so that we can understand how to control the conformance to the assigned technique, and discover characteristic differences between the single PBR techniques.

Throughout the pilot study and the 1995 run we realized that there are some threats to validity. For us it was important to describe and address all of them in detail so that other researchers benefits from the lessons we have learned and can try to avoid the threats while replicating this experiment or developing another one. Some threats have their origin in the fact that this was not an experiment with students but with professionals from industry. Some of the threats might even only be addressed through replication.

We need to replicate the generic part of the experiment in other environments, perhaps even in other countries where differences in language and culture may cause effects that can be interesting targets for further investigation. These replications can take the form of controlled experiments with students, controlled experiments with subjects from the industry using their usual technique for comparison, or case studies in industrial projects.

One challenging goal of a continued series of experiments will be to assess the impact that the threats to validity have. Since it is often hard to design the experiment in a way that controls for most of the threats, a possibility would be to concentrate on certain threats in each replication to assess their impact on the results. For example, one replication may use control groups to measure the effect of repeated tests, while another replication may test explicitly for maturation effects. However, we need to keep the replications under control as far as threats to external validity are concerned, since we need to assume that the effects we observe in one replication will also occur in the others.

We are currently working on a lab package for researchers to support the replication of this experiment by other researchers in other environments.

Acknowledgements

This research was sponsored in part by grant NSG-5123 from NASA Goddard Space Flight Center to the University of Maryland. We would also like to thank the members of the Experimental Software Engineering Group at the University of Maryland for their valuable comments to this paper.

References

(Campbell, 1963) Campbell, D. T. and Stanley, J. C. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston, MA: Houghton Mifflin Company.

(Edington 1987) Edington, E. S. 1987. *Randomization Tests*. New York, NY: Marcel Dekker Inc.

(Fagan, 1976) Fagan, M. E. 1976. *Design and code inspections to reduce errors in program development*. IBM Systems Journal, 15(3):182-211.

(Hatcher, 1994) Hatcher, L. and Stepanski, E. J. 1994. *A Step-by-Step Approach to Using the SAS® System for Univariate and Multivariate Statistics*. Cary, NC: SAS Institute Inc.[8]

(Heninger, 1980) Heninger, K. L. 1985. *Specifying Software Requirements for Complex Systems: New Techniques and Their Application*. IEEE Transaction on Software Engineering, SE-6(1):2-13.

(Humphrey, 1996) Humphrey, W. S. 1996. *Using a Defined and Measured Personal Software Process*. IEEE Software. 13(3): 77-88.

[8] SAS® is the registered trademark of SAS Institute Inc.

(Linger, 1979) Linger, R. C., Mills, H. D. and Witt, B. I. 1979. *Structured Programming: Theory and Practice*. In The Systems Programming Series. Addison Wesley.

(Parnas, 1985) Parnas, D. L. and Weiss, D. M. 1985. *Active design reviews: principles and practices*. In Proceedings of the 8th International Conference on Software Engineering, pp.215-222.

(Porter, 1995) Porter, A. A., Votta, L. G. Jr. and Basili, V. R. 1995. *Comparing Detection Methods For Software Requirements Inspections: A Replicated Experiment*. IEEE Transactions on Software Engineering, 21(6): 563-575.

(SAS, 1989) SAS Institute Inc. 1989. *SAS/STAT User's Guide, Version 6, Fourth edition, Vol.2*. Cary, NC: SAS Institute Inc.[9]

(SEL, 1992) Software Engineering Laboratory Series. 1992. *Recommended Approach to Software Development, Revision 3*. SEL-81-305, pp.41-62.

(Shapiro, 1965) Shapiro, S. S. and Wilk, M. B. 1965. *An Analysis of Variance Test for normality (concrete samples)*. Biometrika, 52: 591-611.

(Votta, 1993) Votta, L. G. Jr. 1993. *Does every inspection need a meeting?*. In Proceedings of ACM SIGSOFT '93 Symposium on Foundations of Software Engineering. Association of Computing Machinery.

(Winer, 1991) Winer, B. J., Brown, D. R. and Michels, K. M. 1991. *Statistical Principles in Experimental Design, 3rd ed*. New York, NY: McGraw-Hill Inc.

[9] JMP® is a trademark of SAS Institute Inc

A. Sample Requirements

Below is a sample requirement from the ATM document which tells what is expected when the bank computer gets a request from the ATM to verify an account:

Functional requirement 1 (From ATM document)

Description: The bank computer checks if the bank code is valid. A bank code is valid if the cash card was issued by the bank.

Input: Request from the ATM to verify card (Serial number and password)

Processing: Check if the cash card was issued by the bank.

Output: Valid or invalid bank code.

We also include a sample requirement from one of the NASA documents in order to give a picture of the difference in nature between the two domains. Below is the process step for calculating adjusted measurement times:

Calculate Adjusted Measurement Times: Process (From NASA document)

1. Compute the adjusted Sun angle time from the new packet by

$$t_{s,adj} = t_s + t_{s,bias}$$

2. Compute the adjusted MTA measurement time from the new packet by

$$t_{T,adj} = t_T + t_{T,bias}$$

3. Compute the adjusted nadir angle time from the new packet.

a. Select the most recent Earth_in crossing time that occurs before the Earth_in crossing time of the new packet. Note that the Earth_in crossing time may be from a previous packet. Check that the times are part of the same spin period by

$$t_{e-in} - t_{e-out} < E_{\max} T_{spin,user}$$

b. If the Earth_in and Earth_out crossing times are part of the same spin period, compute the adjusted nadir angle time by

$$t_{e-adj} = \frac{t_{e-in} + t_{e-out}}{2} + t_{e,bias}$$

4. Add the new packet adjusted times, measurements, and quality flags into the first buffer position, shifting the remainder of the buffer appropriately.

5. The Nth buffer position indicates the current measurements, observation times, and quality flags, to be used in the remaining Adjust Processed Data section. If the Nth buffer does not contain all of the adjusted times (and), set the corresponding time quality flags to indicate invalid data.