The Empirical Investigation of Perspective-Based Reading

Victor R. Basili[1], Scott Green[2], Oliver Laitenberger[3],
Forrest Shull[1], Sivert Sørumgård[4], Marvin V. Zelkowitz[1]

[1] Computer Science Department/
Institute for Advanced Computer Studies
University of Maryland, College Park, MD, 20742
{basili, fshull, mvz}@cs.umd.edu

[2] NASA Goddard Space Flight Center
Code 552.1
Greenbelt, MD, 20771
scott.green@gsfc.nasa.gov

[3] AG Software Engineering
Fachbereich Informatik
Universität Kaiserslautern
Postfach 3049
67653 Kaiserslautern
Germany
laitenbe@informatik.uni-kl.de

[4] The Norwegian Institute of Technology
The University of Trondheim
UNIT/NTH-IDT
O.S. Bragstads plass 2E
Trondheim, N-7034
Norway
sivert@idt.unit.no

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 Perspective-Based Reading (PBR), a particular reading technique for 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 than the same number of readers using their usual technique.

This work was supported in part by NASA grant NSG-5123 and UMIACS.

To test the efficacy of PBR, we conducted two runs of a controlled experiment in the environment of the National Aeronautics and Space Administration / Goddard Space Flight Center (NASA/GSFC) Software Engineering Laboratory (SEL), using developers from the environment. The subjects read two types of documents, one generic in nature and the other from the NASA domain, using two reading techniques, PBR and their usual technique. The results from these experiments, as well as the experimental design, are presented and analyzed. When there is a statistically significant distinction, PBR performs better than the subjects' usual technique. However, PBR appears to be more effective on the generic documents than on the NASA documents.

1. Introduction

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 lifecycle. 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 reading by step-wise abstraction (Linger, 1979) as developed by Harlan Mills, there has been very little research focused on the development of reading techniques. Most efforts have been associated with the methods (e.g., inspections, walk-throughs, reviews) surrounding the reading technique. In general, 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 the area of programming languages, 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 Software Engineering Laboratory (SEL) environment, we have learned much about the efficacy of reading and reading-based approaches through the application and evaluation of methodologies such as Cleanroom. We are now part of a group (ISERN[1]) that has undertaken a research program to define and evaluate software reading techniques to support the various review methods for software development.

In this paper, we use the following convention to differentiate a "technique" from a "method": A technique is a series of steps, producing some desired effect, and requiring skilled application. We define a method as a management procedure for applying techniques.

1.1 Experimental Context: Scenario-Based Reading

In our attempt to define reading techniques, we established several goals:

- 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 an abstraction of the product, and then answer questions based on the abstraction. 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.

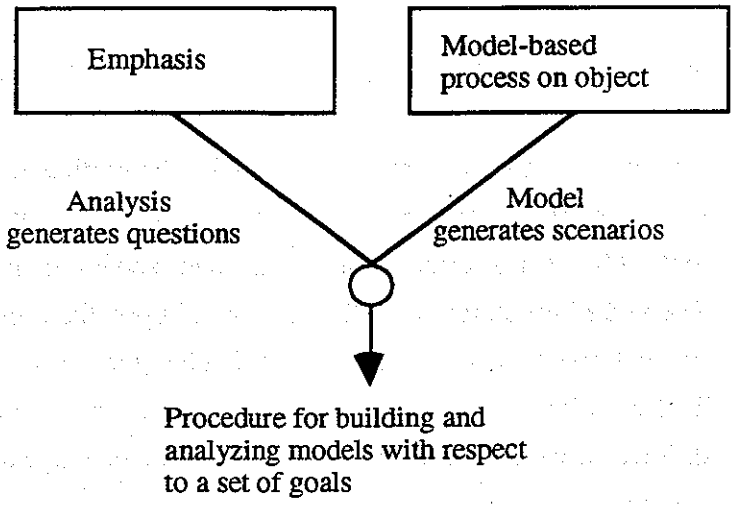

Emphasis
Model-based process on object
Analysis generates questions
Model generates scenarios
Procedure for building and analyzing models with respect to a set of goals

Figure 1. Building focused, tailored reading techniques.

So far, two different 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 in this series. Defect-based reading was defined for reading SCR (Software Cost Reduction) style documents (Heninger, 1980), 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) was undertaken to analyze 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.

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. In this type of scenario-based reading, 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. 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 sufficient coverage of the document but does not cause any particular reader to be responsible for everything.

This work on PBR was conducted within the confines of the NASA/GSFC Software Engineering Laboratory. The SEL, started in 1976, has been developing technology aimed at improving the process of developing flight dynamics software within NASA/GSFC. This class of software is typically written in any of several programming languages, including FORTRAN, C, C++, and 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.

Assume we embed these requirements reading scenarios in a particular method. It then becomes the role of the method to determine which scenarios to apply to the document, how many readers will play each role, etc. This could be done by assuming, as entry criteria, that the method has available to it the anticipated defect class distribution, together with knowledge of the organization's ability to apply certain techniques effectively. Note that embedding focused reading techniques in a method such as inspections provides more meaning to the "team" concept. That is, it gives the readers different views of the document, allowing each of the readers to be responsible for their own view, with the union of the readers providing greater coverage than any of the individual readers.

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

> 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 the test-based reader will use to review the document.

We developed two different series of experiments for evaluating scenario-based techniques. The first series of experiments are aimed at discovering if scenario-based reading is more effective than current practices. This paper's goal is to analyze perspective-based reading and the current NASA SEL reading technique to evaluate and compare them with respect to their effect on fault detection effectiveness. 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 scenario-based reading techniques is most effective.

1.3 Experimental Plan

Our method for evaluating PBR was to see if the approach was more effective than 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 were designed, where 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. We were initially interested in several outcomes:

1. Would individual performances improve if each individual used one of the PBR (designer, tester, user) scenarios in order to find defects?

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

3. Are there characteristic differences in the class of defects each scenario uncovered?

While we were interested in the effectiveness of PBR within our SEL environment, we were also interested in the general applicability of the technique 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 could be reused in other domains.

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 classes of 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 known classes of errors. These applications included an automated parking garage control system, labeled PG, and an automated bank teller machine, labeled ATM.

1.4. Structure of this Paper

In section 2, we discuss how we developed a design for the experiment outlined above. Major issues concerning constraints and threats to validity are described in order to highlight some of the tradeoffs made. We also include a short overview of how the experiment was actually carried out.

Section 3 presents the statistical analysis of the data we obtained in the experiment. The section examines individual results and team results. In each of these parts, we look at the results from both experiment runs, within documents and within domains.

Section 4 is an interpretation of the results of the experiment, but without the rigor of a formal statistical approach. The presentation is again divided into individual results and team results, with concentration on what effect the differences between the two runs of the experiment had in terms of results.

Section 5 summarizes our experiences regarding designing and carrying out the experiment.

2. Design of the Experiment

In this section, we discuss various ways of organizing the individual subjects and the instrumentation of the experiment to test various hypotheses. Two runs of the experiment were conducted. Due to the experiences gained in the initial run, some modifications were introduced in its replication. Differences between the two runs of the experiment will be pointed out where appropriate.

For both experiments, the population was software developers from the NASA SEL environment. The selection of subjects from this sample was not random, since everyone in the population could not be expected to be willing or have opportunity to participate. Thus, all subjects were volunteers, and we accepted everyone who volunteered. Nobody participated in both runs of the experiment.

2.1 Hypotheses

We formulated our main question in the form of the following two hypotheses, where H_0 is the null-hypothesis and H_a is the alternative hypothesis:

H_0 *There is no significant difference in the defect detection rates of teams applying PBR as compared to teams using the usual NASA technique.*

H_a *The defect detection rates of teams applying PBR are significantly higher as compared to teams using the usual NASA technique.*

Our hypotheses are focused on the performance of teams, but we will also analyze the results for the individual performance of the subjects. We make no assumptions at this level regarding the validity of the hypotheses when changing important factors such as subjects, and documents. The constraints relevant for this particular experiment will be explicitly discussed throughout this section, as will the generalizability of the results of the experiment.

2.2 Factors in the Design

In designing the experiment, we had to consider what factors were likely to have an impact on the results. Each of these factors will cause a rival hypothesis to exist in addition to the hypotheses we mentioned previously. The design of the experiment has to take these factors, called *independent variables*, into account and allow 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, which we identify according to how they can be manipulated. Some of them can be controlled during the course of the experiment, while some are strictly functions of time, and still others are not even measurable.

- **Controllable variables:**

  - **Reading technique:** We have two alternatives: One is the technique we have developed, PBR, and the other is the technique currently used for requirements document review in the NASA SEL environment, which we refer to as the "usual" technique.
  - **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.
  - **Perspective:** For PBR, a subject can take one of three perspectives as previously described: Designer, Tester or User.

- **Measurable variables:**

  - **Replication:** This nominal variable is not one we can manipulate, but we need to be aware of its presence because there may be differences in the data from the two experiment runs that may be the result of changes to documents, training sessions or experimental conditions.
  - **Round within the replication:** For each experiment, every subject is involved in a series of treatments and tasks or observations. The results from similar tasks may differ depending on when they take place.

- **Other factors identified:**

  - **Experience:** The experience of each subject is likely to have an impact on the defect detection rate.
  - **Task sequence:** Reading the documents in a sequence may have an influence on the results. This may be a learning effect due to the repetitive reading of multiple documents.
  - **Environment:** The particular environment in which the experiment takes place may have an impact on how well the subjects perform. In this experiment, this effect cannot be separable from effects due to replication.

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 2.5. This section will also cover the last two factors mentioned above: Task Sequence (in the literature referred to as "effects due to testing") and Environment.

2.3 Constraints and Limitations

In designing the experiment we 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:** 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.
- **Subjects:** For the same reasons as stated above, we could not expect to get as many subjects as we would have liked. This required us to be cautious in the design and instrumentation in order to generate as many useful data points as possible.

Specifically, we knew that we could expect to get between 12 and 18 subjects for two days on any run of the experiment.

Another factor that we had to deal with is that we had to provide some potential benefit to the subjects since their organization 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.

2.4 Choosing a Design

Due to the constraints, we found that constructing real teams of (three) reviewers to work together in the experiment would take too much time for the resulting amount of data points. This decision was supported by similar experiments (Parnas, 1985) (Porter, 1995) (Votta, 1993), where the team meetings were reported to have little effect; the meeting gain was outweighed by the meeting loss. However, the team is an important unit in the review process, and PBR is team-oriented in that each reviewer has a responsibility that is not shared by other reviewers on the team. Thus our reviewers did not work together in teams during the course of the experiment. Instead we conducted the experiment based on individual tests, and then used these individual results to construct hypothetical teams after the experiment was completed.

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 possible ways of arranging the order of tasks and treatments for a group of subjects:

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

Option 3, where the subjects first use PBR and then switch to their usual technique, was not considered an 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, but that is a situation we cannot control because the subjects already know their usual technique. Thus, this becomes more a problem in terms of external validity.

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.

- **NASA:**
  - Flight dynamics support (A) - 27 pages, 15 seeded defects
  - Flight dynamics support (B) - 27 pages, 15 seeded defects

- **Training:**
  - Video rental system - 14 pages, 16 seeded defects
  - NASA sample - 9 pages, 6 seeded defects

Since we have sets of different documents and techniques to compare, it became clear that a variant of factorial design would be convenient for this experiment. Such a design would allow us to test the effects of applying both of the techniques on both of the relevant documents. We found that a full factorial design would be inappropriate for two reasons. First, a full factorial design would require some subjects to apply the ordering of techniques that we previously argued against. Secondly, such a design seemed hard to conduct because it would require each subject to use all three perspectives at some point. 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.

The use of control groups to assess differences in documents and learning effect appeared to bear an unreasonable cost, since the use of such groups would decrease the remaining number of data points available for analyzing the difference between the techniques. The low number of data points might result in data that would be heavily biased due to individual differences in performance. Based on the cost and the fact that previous related experiments (Porter, 1995) showed that effects of learning were not significant, we chose not to use control groups. This decision also made the experiment more attractive in terms of getting subjects, since they would all receive the same amount and kind of training.

	Group 1	Group 2	
	D T U	D T U	
NASA technique	Training	Training	First day
	NASA A	NASA B	
	Training	Training	
	ATM	PG	
	Teaching of PBR		
PBR technique	Training	Training	Second day
	PG	ATM	
	Training	Training	
	NASA B	NASA A	

Figure 2. Design of the experiment.

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). The number of subjects was about the same for the two experiments (12-14 subjects).

2.5 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 the experimental design is to minimize the impact of these threats.

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 correlation between 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.

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

- **History:** We need to consider what the subjects did between the pretests and posttests. In addition to receiving a treatment where they were taught a new reading technique, there may have been other events outside of our control that had an impact on the results. 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.
- **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 generous breaks between sessions to suppress this 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. 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, as this would presumably increase the learning effect. Related experiments have reported that effects due to repeated testing are not significant (Porter, 1995).
- **Instrumentation:** These effects are basically due to differences in the way of measuring scores. Our scores were measured by two people independently, and then discussed in order to resolve any disagreement consistently. Thus this effect is not relevant to us.
- **Selection:** Subjects may be assigned to their treatment groups in various ways. In our case there was a difference between the two experiment runs. In the first one, the subjects were assigned roles for PBR based on their normal work in the NASA environment in order to match roles as closely as possible. This was only minimally successful since the sample was not an even mix of people representing the various roles. However, for the replication, the subjects were randomized. Thus effects due to selection may be somewhat relevant for the first experiment, but not for the replication. Since PBR assumes the reviewers in a team use their usual perspectives, the random assignment used in the experiment would presumably lead to an underestimation of the improvement caused by PBR.

Another threat to 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.

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.
- **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 effects:** These effects are due to the experimental environment. Here we have a difference between the two runs of the experiment. In the initial run, the testing was done in the subjects' usual work environment. The subjects received their training in groups, and then returned to their own workspace for the test. For the replication, the experiment was conducted in an artificial setting away from the work environment, similar to a classroom exercise. This may influence the external validity of the experiment, since a non-experimental environment may cause different results.

There are also a number of other possible but minor threats. One of these is the fact that the subjects knew they were part of an experiment. They knew that the purpose of the experiment was to compare reading techniques, and they probably were able to surmise our expectations with respect to the results even if not stated explicitly. However, these aspects are difficult to eliminate in experiments where subjects are trained in one technique while the comparison technique is assumed to be known in advance. A design where they receive equal training in two techniques would be more likely to hide these effects.

2.6 Preparation and Conduction

We wanted the two experiment runs to be as similar as possible in order to avoid difficulties in combining the resulting data, but some changes between the runs were still necessary. We began preparing for the second run by reviewing all documents and forms in order to improve them from an experimental viewpoint. We had some comments from the first experiment run that were helpful in this process. The changes were minor, and most were directed towards language improvement. We changed the seeded defects in three places in one of the generic documents due to a refined and deeper insight into what we would consider a defect. There were some changes to the forms, scenarios and defect classification as well, but again the changes were made to make the documents easier to use and understand.

For the NASA documents, the changes were more fundamental. For the first experiment run, comments from the participants indicated that the documents were too large and complex. We decided to make them shorter and simpler for the second experiment run. As a side effect of this change, the total number of defects in the NASA documents was reduced. However, the types and distribution of seeded defects remained similar.

The basic schedule for conducting the experiment remained unchanged. Each experiment run lasted for two whole work days, with one day off in-between. The number and order of document reviews were also the same for both experiments, but the time allowed for each review was modified. For the first experiment run, the maximum time for one document was three hours. However, for the generic documents, only one person used more than two hours (140 minutes), so under the more controlled environment of the second experiment run, we felt safe lowering the maximum time to two hours.

Another important change resulted from the comments we received from the first experiment run, regarding the training sessions. 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 the second experiment run, we had training sessions before each document review. For this purpose we generated two sample documents that were representative of the NASA and generic domains.

After the second run of the experiment, we marked all reviews with respect to their defect detection rate. This was measured as the percentage of the seeded defects that was found by each reviewer. We did not consider any other measures such as false positives. Based on the defects found by the reviewers, we also refined our understanding of the defects present in the set of documents. 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 first experiment, we re-marked all the reviews from the first experiment in order to make all results consistent.

3. Statistical Analysis

We ran the experiment twice, in November 1994 (hereafter referred to as the "1994 experiment") and in June 1995 (hereafter referred to as the "1995 experiment"). In the 1994 experiment, we had twelve subjects read each document, six using the usual technique and six using PBR. The six using PBR were distributed equally among the three perspectives. In the 1995 experiment, we had thirteen subjects who read each document, although a fourteenth volunteer unfamiliar with the NASA domain also read the generic documents only.

After the two experiment runs, 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.

We attempted to measure 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 several roles (analyst, tester, designer, manager). However, for any realistic measurement scale, most reviewers tended to clump together toward the middle of the range, with relatively few outliers in either direction. Thus we seem to have a relatively homogeneous sample with respect to this variable. While good from an experimental viewpoint, this unfortunately means that our data set does not allow for a meaningful test of the effect of reviewer experience, and we are forced to defer an investigation of the interaction between reading technique and experience until such time as we can get more data points. For this reason, reviewer experience will not appear as a potential effect in any of our analysis models.

Technique, experiment run, and document are represented by nominal-scale variables used in our models, where appropriate. The domain is taken into account by performing a separate analysis for each of the generic and NASA problem domains. However, 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 2.5, and any interpretation of results must take them into account.

Section 3.1 presents the details of the effect on individual scores. Section 3.2 presents the analysis strategy for team data. Finally, Section 3.3 takes an initial look at the analysis with respect to the reviewer perspectives. In each section, we present the general analysis strategy and some details on the statistical tests, followed by the statistical results and some interpretation of their meaning. We address the significance of our results taken as a whole in Section 4.

3.1 Analysis for Individuals

Although it was not part of our main hypothesis, which focuses on team coverage, we wanted to see if the difference in focus between the usual technique and PBR would have some effect on individual detection rates. We therefore went through an analysis of individual scores.

We were also careful, however, to test for effects from sources of variation other than the reading technique. For this reason, our analysis proceeds in a "bottom-up" manner. That is, we begin with several small data sets that we know to be homogeneous. Each session of the experiment was run under controlled conditions to eliminate differences within the sessions that might have an effect on reviewers' detection rates; the scores of reviewers reading the same document within the same replication are therefore comparable. Thus we begin our analysis with homogeneous data sets (4 documents - 2 NASA and 2 generic - over 2 runs, so 8 in total) which we will use as the primary building blocks of our analysis.

Starting from these data sets, we looked for features in common between the data sets. We identified subsets of the data which were expected to be more homogeneous than the data as a whole; the aim was to exploit this homogeneity to achieve stronger statistical results. For example, we took into account the fact that all of the detection rates for each reviewer are highly correlated, but we also identified other such blocks (e.g., the data for each problem domain within the experiment). As we looked at larger data sets in order to draw more general conclusions, we also took pains to make sure that the data within each set were still comparable. Figure 3 illustrates the direction of our analysis, and includes the sizes of the data sets.

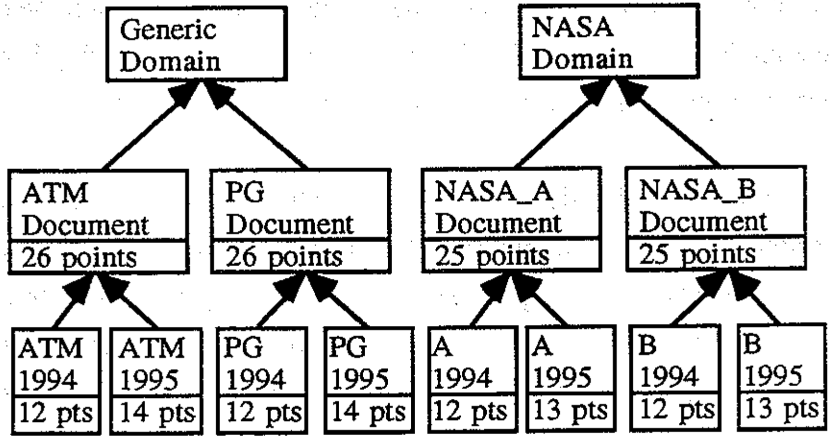


Figure 3. Breakdown of the statistical analysis, with number of data points.

3.1.1 Analysis Strategy Within Documents

Our initial analysis examined each document used in the experiment for significant differences in performance based on the use of reading technique. We used the ANOVA test since we were testing a model of the effects containing multiple potential sources of variation. To begin with, our model of the effects contained a nominal variable to signify the reading technique used (usual or PBR).

The data for each document is composed of the independent data sets from the two experiment runs, and so it was necessary to be alert to the possibility that changes from one run to the next could have an impact on the reviewers' detection rates. For both of the pairs of documents, we combined the data for the document and introduced a nominal variable (with two levels: 1994 and 1995) into our model to describe the experiment run in which the reviewer read the document.

We measured the lack of fit error (an estimate of the error variance) for the model on each document. In no case was there a significant lack of fit error, so it did not seem likely that we could gain any better fit to the data by introducing variations on the variables, such as testing for interaction effects (SAS, 1989).

We also tested whether each of the variables independently was significant (i.e., whether the effect of each variable, apart from the other variables in the model, had a significant effect on reviewer detection rate).

The ANOVA test 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). Also, the test assumes that variance between samples for each level of the independent variables is homogeneous. 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 2.5, "Selection" and "Interaction of selection and treatment").

3.1.2 Results Within Documents

In our case the hypotheses of the ANOVA test take the following form:

H_0: The specified model (which contains variables to signify the experiment run and reading technique) has no significant power in predicting the value of the dependent variable (detection rate).

H_a: The model as a whole is a significant predictor of detection rate.

Level of significance: $\alpha = 0.05$

The ANOVA test also allows testing the effect of each individual variable.

The Least Squares Means (LSM) of the detection rates for reviewers using each of the techniques are given in Table 1, followed by the results of the tests for significance. The LSM values in effect allow an examination of the means for the groups using each of the reading techniques while holding the difference due to experiment run constant. This is followed by the p-values resulting from the statistical tests for significance; a p-value of less than 0.05 provides evidence that either the whole model or the individual variable is a significant predictor of detection rate and are indicated in boldface. The R^2 value for the model is also included as a measure of the amount of variation in the data that is accounted for by the model.

For all documents except NASA_B, the LSM detection rate for PBR reviewers is slightly higher than for reviewers using their usual technique. However, only for the ATM document was the difference statistically significant. For all other documents, reviewers using the two techniques did roughly the same, and any differences between their average scores can be attributed to random effects alone. Both NASA documents had a very significant effect due to experiment run, which was not surprising, given the large changes made to improve the documents between runs; however, there was also a significant and unexpected effect due to experiment run for the PG document as well. The significance of such differences due to experiment run is addressed in Section 4.

Document	PBR LSM	USUAL LSM	Whole Model p-value	Technique p-value	Replication p-value	R^2
ATM	30.8	21.4	0.0904	**0.0316**	0.6299	0.19
PG	26.8	24.5	**0.0457**	0.5977	**0.0174**	0.24
NASA_A	36.8	26.6	**0.0001**	0.1516	**0.0001**	0.56
NASA_B	28.3	34.5	**0.0021**	0.5044	**0.0005**	0.43

Table 1. Effects on individual scores for each document.

3.1.3 Analysis Strategy Within Domains

The second level of detail which we analyzed was the level of problem domains. That is, we examined what trends could be observed within the generic documents or within the NASA documents, while realizing that such trends may not necessarily apply across such different domains. For each domain, we tested whether each reviewer scored about the same when reviewing documents with PBR as when using the usual technique, or if there was in fact a significant effect due to reading technique.

To accomplish this, we made use of the MANOVA (Multivariate ANOVA) test with repeated measures, an extension of the ANOVA which measures effects across multiple dependent variables (here, the scores on each of the two documents) with longitudinal data sets (i.e. data sets in which each subject is represented by multiple data points).

The domain data sets contain two scores for each subject, one for each document within the domain. Although repeated measures tests usually refer to multiple treatments over time, here we treat the scores on each document as the scores from repeated treatments, which we distinguish with the nominal variable "Document". We divide the reviewers into two groups, and use another nominal variable in order to distinguish to which group each reviewer belonged: Group I applied PBR to Document A and the usual technique to Document B, and Group II read the documents in the opposite fashion. If the interaction between these two variables is significant, we can conclude that the reading technique a reviewer applied to each document had a significant effect on the reviewer's detection rate. If the interaction is not significant, then reviewers tended to perform about the same on the two documents, regardless of the technique applied to each. Aside from reading technique and document, we again want to account for any significant effects due to the experiment run, and also test for interaction effects between this variable and the others.

The MANOVA test with Repeated Measures makes certain assumptions about the data set. As with the ANOVA test, we again fulfill requirements about the measurement scales of the dependent and independent variables, the independence of observations, and the underlying distribution of the sample. We have the same threat to validity resulting from the assumption of random samples as was discussed for the ANOVA test. However, it is also assumed that the dependent-variable covariance matrix for a given treatment group should be equal to the covariance matrix for each of the remaining groups. Fortunately, the type I error rate is relatively robust against typical violations of this assumption; however, the power of the test is somewhat attenuated (Hatcher, 1994).

3.1.4 Results Within Domains

Using the data from each of the documents within a domain, we use the MANOVA test to detect how reviewer rates change from one document to the next, and attribute these changes to factors in our model. As we did with the ANOVA test, we test whether each of the variables in our model (the documents themselves, the reading technique used on each document, the experiment run, and all appropriate interactions) are significant predictors of the change in detection rates.

H_0: The specified variable has no significant effect in predicting scores across the two documents.

H_a: The variable is a significant predictor of scores across the documents.

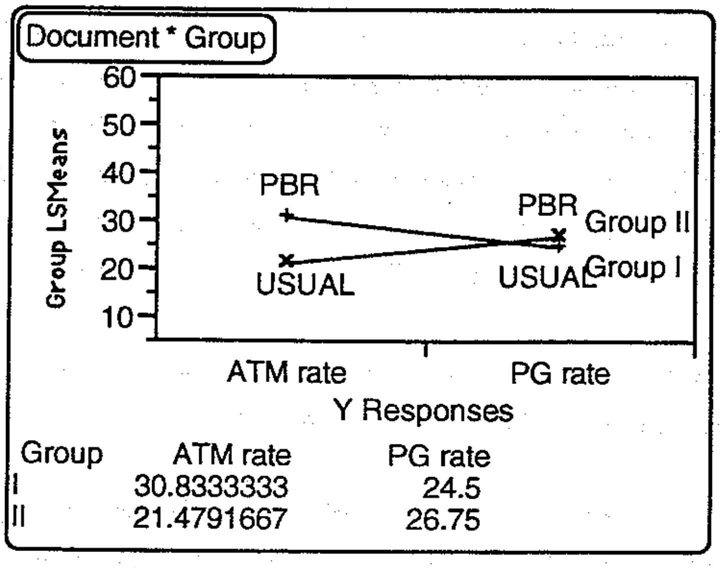
Level of significance: $\alpha = 0.05$

The results are summarized in Table 2, where each column gives the p-value for each of the effects. A p-value of less than 0.05 provides an indication that the variable is a significant predictor of the change in reviewer detection rates across documents, and appears in bold. The effect due to the reading technique is measured indirectly by the "Group" variable: Group I read Document A with the PBR technique and Document B with the usual technique; Group II read the documents in the reverse fashion. As can be seen from the "Document" column, there was no significant difference between the mean detection rates for the two documents within a domain. Crossed terms represent tests for interaction effects; for example, the column labeled "Document * Replication" tests if the mean difference in reviewers' scores on each of the documents was significantly effected by the experiment run in which they took part. Thus, even though the NASA documents were changed drastically between runs, because the two documents were roughly comparable in difficulty within both experiment runs, there is no significant effect here for the NASA domain. Within the generic domain, reviewers in the 1994 experiment did slightly better on the PG document than the ATM, while reviewers in the 1995 experiment did slightly worse on the PG document relative to the ATM; while the differences average out when the two runs are combined, the effect still shows up as a significant interaction in the MANOVA test.

Domain	Document	Document * Replication	Document * Group	Document * Replication * Group
Generic	0.7810	**0.0298**	**0.0056**	0.5252
NASA	0.9137	0.7672	0.5670	0.5337

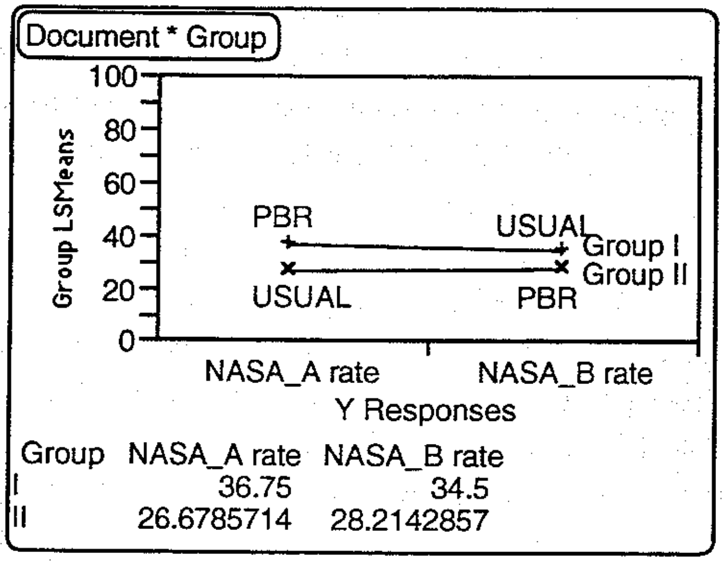
Table 2. Effects on individual scores within domains.

Graphs of Least Squares Means are presented in figures 4a and 4b as a convenient way of visualizing the effects of the interaction between document and reading technique. For the generic domain, it can be seen that reviewers in each group on average scored higher with PBR than with the usual technique, taking into account the other effects in the model. In the NASA domain, reviewers in each group scored about the same on both documents, regardless of the technique used. Note that the interaction for the generic domain is significant, providing evidence that reading technique does in fact have an impact on detection rates.



Group	ATM rate	PG rate
I	30.8333333	24.5
II	21.4791667	26.75

Figure 4a. Interaction between group and technique for the generic domain.



Group	NASA_A rate	NASA_B rate
I	36.75	34.5
II	26.6785714	28.2142857

Figure 4b. Interaction between group and technique for the NASA domain.

3.2 Analysis for Teams

3.2.1 Analysis Strategy for Teams

In this section, we return to investigating our primary hypothesis concerning the effect of PBR on inspection teams. The analysis was complicated by the fact that the teams were composed after the experiment's conclusion, and so any grouping of individual reviewers into a team is somewhat arbitrary, 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 techniques. An informal description of the test follows.

First, since there are differences between the experiment runs, we will compose teams only with reviewers from within the same run; we therefore treat the two experiment runs separately. Results from the individual scores showed that the domains are very different, but the documents within a domain are of comparable difficulty; thus, we compare reviewer scores on documents within the same domain only. We again categorize reviewers into one of two groups, as we did for the analysis within domains for individual scores, depending on which technique they applied to which document. Let us say the reviewers in Group I applied PBR to Document A and their usual technique to Document B, where Document A and Document B represent the two documents within either of the domains. We can then generate all possible PBR teams for Document A and all possible non-PBR teams for Document B, and take 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_I and B_I. We can then perform the same calculations for Group II, in which reviewers applied their usual technique to Document A and PBR to Document B, in order to obtain averages A_{II} and B_{II}. The test statistic

$$(A_I - B_I) - (A_{II} - B_{II})$$

then gives us some measure of how all possible PBR teams would have performed relative to all possible usual technique teams.

Now suppose we switch a reviewer in Group I with someone from Group II. The new reviewer in Group I will be part of a PBR team for document A even though he used the usual technique on this document, and will be part of a usual technique team for Document B even though he applied PBR. A similar but reversed situation awaits the reviewer who suddenly finds himself in Group II. 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, the average score will decrease, even as the average score of usual technique teams with more and more PBR members is being raised. 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 I and Group II, and rank each of these scenarios in decreasing order by the statistic. If the scenario in which no dilution has occurred appears toward the top of the list (in the top 5%) we will conclude PBR does have a beneficial effect on team scores, since every time the PBR teams were diluted with non-PBR reviewers they tended to perform somewhat worse relative to the usual technique teams. However, should the non-diluted scenario appear toward the middle of the list, then this is clear evidence that every successive dilution had only random effects on team scores, and thus that reading technique is not correlated with team performance.

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).

3.2.2 Results for Teams

The use of the permutation test allows us to formulate and test the following hypotheses:

H_0: The difference between average scores for PBR and usual technique teams is the same for any random assignment of reviewers to groups.

H_a: The difference between average scores for PBR and usual technique teams is significantly higher when the PBR teams are composed of only PBR reviewers and the usual technique teams are composed of only usual technique reviewers.

Level of significance: $\alpha = 0.05$ (that is, we reject H_0 if the undiluted teams appear in the top 5% of all possible permutations between groups)

The results are summarized in Table 3. P-values which are significant at the 0.05-level appear in boldface. For example, twelve reviewers read the generic documents in the 1994 experiment; 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.

Domain/ Replication	**Number of Group Permutations Generated**	**Rank of Undiluted Group**	**P-value**
Generics/1995	3003	2	**0.0007**
Generics/1994	924	61	0.0660
NASA/1995	1716	67	**0.0390**
NASA/1994	924	401	0.4340

Table 3. Results of permutation tests for team scores.

3.3 Analysis for Perspectives

3.3.1 Analysis Strategy for Perspectives

We were also concerned with the question of whether the perspectives used in the experiment are useful (i.e., reviewers using each perspective contributed a significant share of the total defects detected) and 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.

3.3.2 Results for Perspectives

We formulate no explicit statistical tests concerning the detection rates of reviewers using each of the perspectives, but present Figures 5a and 5b 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 NASA_A charts as an example from the NASA domain. However, due to the differences between experiment runs for the NASA documents, we do not present a coverage diagram for both runs combined. 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 experiment 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.

ATM Results:

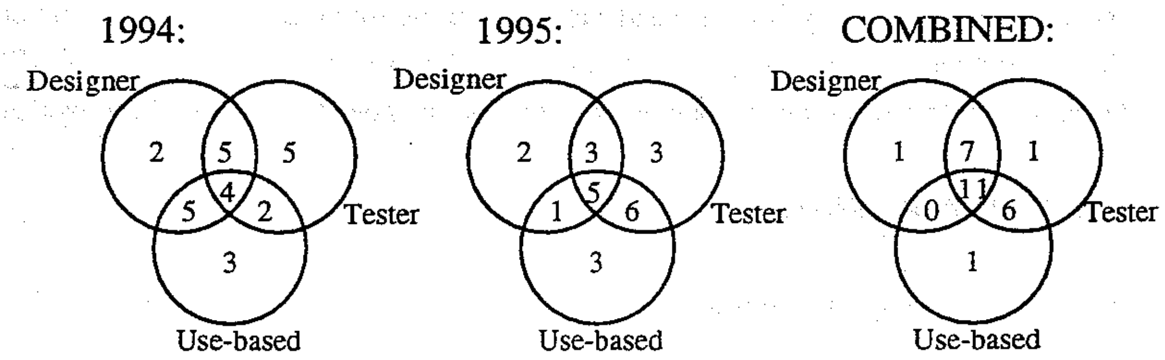


Figure 5a. Defect coverage for the ATM document.

NASA_A Results:

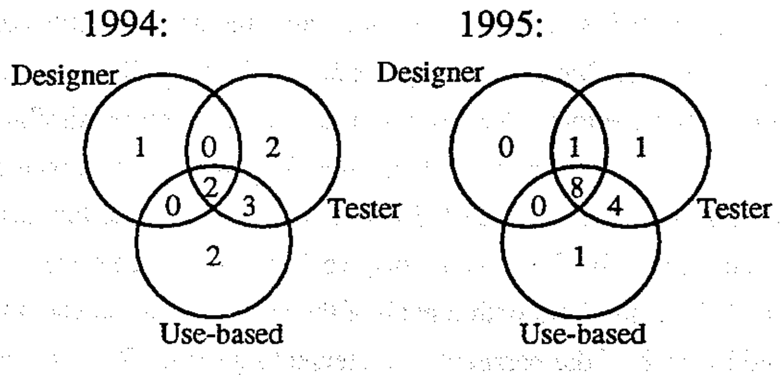


Figure 5b. Defect coverage for the NASA_A document.

4. PBR Effectiveness

In the previous section we presented the analysis of the data from a strictly statistical point of view. However, it is necessary to assess the meaning and implications of the analysis to see if we can identify trends in the results that are similar for both runs of the experiment. Such interpretations may also point out areas of weakness in the experiment or in the PBR technique - weaknesses which upon recognition become potential areas for improvement.

4.1. Individual Effectiveness

4.1.1. The 1994 Experiment

The individual defect detection rates were better for the generic documents than for the NASA documents in the 1994 replication, regardless of reading technique, because the generic documents were simpler to read and less complex than the NASA documents. Most subjects pointed to the size and complexity of the NASA documents as potential problem areas. However, there is a difference not only in absolute score, but also in the impact the technique has on detection rate. The improvement of PBR over the usual technique was greater for the generic documents than for the NASA documents. We can think of various reasons for this:

- The perspectives and the questions provided were not aimed specifically at the NASA documents, but based on the general nature of the generic documents.

Thus the technique itself may not be exploited to its full potential for documents within the NASA domain.

- It is possible that the reviewers are more likely to fall back on their usual technique rather than apply the PBR technique when reading documents that they are familiar with. We received anecdotal evidence of this during follow-up interviews. This may be of particular importance in situations where the subjects are under pressure due to time constraints and the complexity of the document.
- The 1994 experiment was carried out in the reviewers' own work environment. This may increase the temptation to fall back to the usual technique when the familiar situation of reading NASA documents arose. The generic documents, on the other hand, would not be likely to stimulate such interaction effects.
- Insufficient training may have been provided since the training sessions only explained how to use the technique on a sample generic document and not on a sample NASA document.

Within each of the two domains, we found that the documents were at the same level of complexity with only minor differences between them. This indicated that our effort of keeping the documents within each domain comparable was successful.

4.1.2. The 1995 Experiment

In the 1995 replication we made some changes to account for some of the problems mentioned above. The NASA documents were modified substantially according to the comments we received from the subjects. We also provided additional training by adding two more sessions aimed at applying the techniques to the NASA documents. The experiment itself was carried out in a classroom environment instead of the work environment. However, even though we saw a substantial rise in the absolute defect detection rates for the NASA documents, the improvement of PBR over the usual technique remained insignificant. Thus our most viable explanation at the moment is that PBR needs to be more carefully tailored to the specific characteristics of the NASA documents and environment to show an improvement similar to what we see in the generic domain. 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.

For the generic domain, we made only minor changes to the documents and the seeded defects. Thus, we expected the change in defect detection rate to be negligible. However, this appeared not to be the case.

The mean detection rate for the ATM document turned out to remain unchanged, but dropped significantly for the PG document. We have analyzed this carefully, but have not been able to find a plausible explanation as to why this should happen. Changes to the experiment should be expected to have a similar impact on the two documents, so perhaps the changes to the two documents were not as insignificant as we thought.

4.1.3. Combined

Although the changes to the NASA documents were a definite improvement, any effect due to technique is hidden by the much larger difference between the two runs of the experiment. This problem illustrates one of the tradeoffs we had to make when planning the second run. Should we have kept the documents unchanged, thus getting data that may not be completely valid, or should we change the documents but get data that would be hard to combine with the data from the initial run? We chose to change the documents, and in retrospect we feel the right decision was made.

We did not have the same problems with the generic documents because they were changed only slightly between the two runs of the experiment. Thus the results indicate a significant improvement of the defect detection rate in the generic domain due to the application of PBR.

4.2. Teams

4.2.1. The 1994 Experiment

The defect detection rates of teams in the 1994 experiment reflected the same trends as the individual rates. For the NASA documents, the defect detection rates were much lower than they were for the generic documents, regardless of reading technique. But even more importantly, the results from the permutation test indicate that there are only random differences between the two techniques in this case. This, together with the defect coverage discussed in section 3.2, counts as evidence that the current perspectives do not work as well with the NASA documents as they do with the generic documents.

4.2.2. The 1995 Experiment

In the 1995 experiment, the team results for the generic documents showed that using PBR resulted in a significant improvement over the usual technique. The reasons for this observed improvement, as compared to the 1994 experiment, may include better training sessions and a less intrusive environment, which in the 1995 experiment was a classroom setting. This environment may have made it easier to concentrate on the experiment and thus to keep the two techniques independent from each other.

For the NASA documents, the results were also better than in 1994. In addition to the possible explanations mentioned for the generic documents, there is the fact that there were substantial changes to the documents. Thus, the results provide more evidence for the 1994 indication that the subjects tend to use their usual technique when reading familiar documents in a familiar work environment, and in particular when under pressure.

4.3. Threats to Validity

The threats to internal validity discussed in section 2 may have an impact on the results of the experiment. Thus, at this point it may be interesting to see whether the potential impact and the results agree. Below we discuss the threats that we find most important:

- **History:** One problem with our experiment is that it does not allow history effects to be separated from the change in technique. 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. We do not consider this effect to be very significant, but we cannot completely ignore it.
- **Maturation:** We may assume the results obtained in the afternoon to be worse than the results from the morning session because the subjects may get tired and bored. 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.

- **Testing:** This may result in an improvement in defect detection rate due to learning the techniques, becoming familiar with the documents, becoming used to the experimental environment and the tests. 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.
- **Reactive effects:** The change of experimental environment between the experiment runs may have made it easier to concentrate on the techniques and tests to be done, thus separating the techniques better for the second run of the experiment.

We cannot say anything conclusive about the impact of threats to validity. However, we feel that we have taken them into account as carefully as possible, given the nature of the problem and our experimental design.

Since the two runs of this experiment have been done in close cooperation with the NASA SEL environment, it seems natural to conclude this section with a discussion of the extent to which 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 experiments 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. Finally, there is a potential threat to validity in the choice of experimental environment. In 1994, the experiment 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.

5. Observations on Experimental Design

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?*

There appears to be no easy answer to this question. Each design will be a result of a number of tradeoffs, and it is not always possible to know how the decisions will influence the data. A good design can have various interpretations based on what are considered the goals for the experiment. One option is to use different designs involving different threats to validity and study the results as a whole.

- *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. For this experiment, an example of such a tradeoff is that we chose to neglect learning effects in order to avoid spending subjects on control groups. This gave us more data points to be used in analyzing the difference between the two techniques, but at the same time 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?*

  Our problem as experimenters is to maintain a certain level of validity while still generating sufficient interest for an organization to allow us to conduct the experiment. From an organization's point of view, an experiment should be closely tied to their own environment to see if the suggested improvement works with minimal effort in terms of environmental changes. From an experimental point of view, however, we are interested in a controlled environment where disturbing interaction effects are negligible.

- *To what extent can experimental aspects such as design, instrumentation and environment be changed when the experiment still is to be considered a replication?*

  One requirement for being considered a replication is that the main hypotheses are the same. Changes in design and instrumentation, in particular to overcome threats to validity, should also be considered "legal". However, one situation we should avoid is making substantial changes to the design based on the *results* from a previous experiment. This will introduce dependencies between the experiments that are highly undesirable from a statistical point of view.

For this experiment in particular, there are various problems that we need to study more carefully. The threats to validity should be carefully examined; in particular we feel the testing effects to be crucial. An experiment with a control group could be one way of estimating what the importance of these effects really are. We may also consider a more careful analysis of the NASA documents and environment in order to refine PBR to these particular needs. The results indicate that the choice of perspectives and associated scenarios do not match the needs of the NASA domain.

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.").

There are numerous alternative directions for the continuation of this research. For further experimentation within NASA's SEL it seems to be necessary to tailor PBR to more closely match the particular needs of that domain. A possible way of further experimentation would be to do a case-study of a NASA SEL project to obtain more qualitative data.

We may also consider replication of 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.

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, Donald T. and Stanley, Julian C. 1963. *Experimental and Quasi-Experimental Designs for Research* . Boston, MA: Houghton Mifflin Company.

(Edington 1987) Edington, Eugene 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, Larry and Stepanski, Edward J. 1994. *A Step-by-Step Approach to Using the SAS® System for Univariate and Multivariate Statistics*. Cary, NC: SAS Institute Inc.[2]

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

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

(Porter, 1995) Porter, Adam A., Votta, Lawrence G. Jr. and Basili, Victor R. *Comparing Detection Methods For Software Requirements Inspections: A Replicated Experiment.* IEEE Transactions on Software Engineering, June 1995.

(SAS, 1989) SAS Institute Inc. 1989. *JMP® User's Guide.* Cary, NC: SAS Institute Inc.[3]

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

(Votta, 1993) Votta, Lawrence 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, December 1993.

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

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

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 ($t_{s,adj}, t_{b,adj}, t_{T,adj}$, and $t_{e,adj}$), set the corresponding time quality flags to indicate invalid data.

Footnotes

[1] ISERN is the International Software Engineering Research Network whose goal is to support experimental research and the replication of experiments.

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

[3] JMP® is a trademark of SAS Institute Inc.