Assessing Defect Detection Methods for Software Requirements Inspections Through External Replication

Filippo Lanubile*, and Giuseppe Visaggio

University of Bari, Dipartimento di Informatica
Via Orabona 4, 70126 Bari, Italy
email: [lanubile|visaggio]@seldi.uniba.it

Abstract

This paper presents the external replication of a controlled experiment which compared three defect detection techniques (Ad Hoc, Checklist, and Defect-based Scenario) for software requirements inspections, and evaluated the benefits of collection meetings after individual reviews. The results of our replication were partially different from those of the original experiment. Unlike the original experiment, we did not find any empirical evidence of better performance when using scenarios. To explain these negative findings we provide a list of hypotheses. On the other hand, the replication confirmed one result of the original experiment: the defect detection rate is not improved by the collection meetings.

The external replication was made possible by the existence of an experimental kit provided by the original investigators. We discuss what difficulties we encountered in applying the package to our environment, having different cultures and skills. We also discovered some critical problems in the original experiment which can be considered threats to its internal validity. Using our results, experience and suggestions, other researchers will be able to improve the original experimental design before attempting further replications.

1. Introduction

Software inspection is the best known way of detecting defects in software requirements specifications (SRS). Initially created for source code (Fagan, 1976), software inspections have been extended for the intermediate products of the earlier phases of the software life cycle, such as design and requirements specifications (Humphrey, 1989).

The basic software inspection process is made up of three steps: detection, collection, and repair. In the detection step, each member of the inspection team individually reviews the document. In the collection step, the members of the team meet together, discussing the results from the individual reviews and preparing a team defect report. Finally, in the repair step, the author of the document fixes the defects contained in the team defect report. This research addresses issues related to the detection and collection steps only.

* Filippo Lanubile is spending a sabbatical period at the University of Maryland, College Park

1.1 Detection techniques

Ad Hoc and Checklist are the most popular defect detection techniques. With the Ad Hoc technique, no guidance is provided during the inspection, but the reviewers use their own knowledge and experience to identify defects in the document. With the Checklist technique, reviewers must answer a list of questions which capture the knowledge from previous inspections. Questions can be general or specific for some classes of defects. The purpose of the questions is to provide some direction to reviewers in looking for defects. Both the Ad Hoc and Checklist approaches are nonsystematic, i.e. there is not a defined and clear procedure to follow. However, on an ordinal scale, Checklist can be considered more systematic than Ad Hoc. Another dimension to characterize these approaches is the reviewer's responsibility; with both Ad Hoc and Checklist approaches, reviewers have general (i.e. they look for all kinds of defects) and identical (i.e. no division of work within a team) responsibilities. Parnas and Weiss (Parnas, 1985) criticized the overlapping of responsibilities and proposed active design reviews, a new form of inspection process where the reviewers had specific responsibilities defined by their use of different checklists.

Based on the same concern about the separation of responsibilities, the Scenario approach has been proposed (Porter, 1995) as a systematic defect detection technique with a specific and distinct responsibility for each reviewer. Defect-based scenarios are a set of procedures for detecting particular classes of defects. A defect-based scenario requires the reviewer to create a model for a specific class of defects (e.g. missing functionality) and then answer a list of model-based questions. Each reviewer follows a scenario for a specific defect class and a team combines distinct scenarios.

Running a controlled experiment, Porter, Votta, and Basili (Porter, 1995) found that:

- the defect detection rate increased about 35% using a Scenario approach as compared to Ad Hoc and Checklist;
- Ad Hoc and Checklist techniques were equivalent with respect to the performance of reviewers;
- reviewers using scenarios found more defects in their target defect class as compared to reviewers using Ad Hoc or Checklist;
- reviewers using scenarios were no less effective at finding defects in other classes than reviewers using Ad Hoc or Checklist;

1.2 Collection meetings

As regards the collection step of the inspection process, a common assumption is that meetings allow the reviewers to detect more defects than individual reading only. Thus meetings, although expensive, are worthwhile. However, an unexpected result from the experiment of Porter, Votta, and Basili put in doubt the necessity of collection meetings, i.e. meetings whose main goal is to collect and validate the defects that have been discovered through a preliminary individual reading. They found that the number of new defects discovered during a collection meeting (meeting gain) is equivalent to the number of true defects

detected by individual inspection but not included in the team defect report compiled during the meeting (meeting loss).

1.3 Motivation

We found these experimental results, and their implications on the inspection process, very interesting. However, since it is not possible to draw final conclusions from a single experiment, we conducted a replication of the experiment of Porter, Votta, and Basili.

A comprehensive definition of replication is in (Judd, 1991):

> "Replication means that other researchers in other settings with different samples attempt to reproduce the research as closely as possible. If the results of the replication are consistent with the original research, we have increased confidence in the hypothesis that the original study supported."

Software engineering, as a scientific discipline, needs research whose primary purpose is replication. Such research is especially concerned with external validity, i.e. the extent to which we can generalize the results to the population of interest in the hypothesis. Frequently, in software engineering research we are not able, or it is not practical, to use random samples from a population in order to increase our ability to generalize. Generalization must then be done by running multiple experiments in different settings and times. However, internal replications (i.e. replications conducted by the same researchers) are not sufficient because the empirical observations in support of a hypothesis may be in error or biased by the original researchers. A scientific hypothesis gains increasing acceptance when external replications (i.e. replications independently conducted by different researchers) arrive at the same conclusions.

Our strict replication of the Porter, Votta, and Basili experiment was made possible by the availability of the experimental material, prepared by the original experimenters in the form of a laboratory package.

2. The Replication Study

Our main interest was in replicating the original research as closely as possible. We decided to make only those minimal changes that were necessary to adapt the first experiment to our environment or to get additional useful information. In the following we describe similarities and differences between our replication and the original experiment, and discuss the problems encountered during the preparation, execution and analysis of the experiment.

2.1 Goals and Hypotheses

The main goal of the original experiment can be defined using a Goal/Question/Metric (GQM) template (Basili, 1994):

Analyze the detection techniques for SRS inspections **for the purpose of** assessment **with respect to** the number and type of defects uncovered **from the point of view of** the researcher

We established hypotheses in order to confirm the results of the original experiment. The original investigators found that systematic approaches with specific, coordinated responsibilities, such as Defect-based Scenario, increase the overall effectiveness of the inspection process. Specifically, our hypotheses were:

- the teams using scenarios would find more defects than teams not using them
- the individuals using a scenario for a specific class of defects would find more defects of that target class as compared to individuals using other techniques
- the individuals using a scenario would find the same number of defects not covered by the scenario as compared to individuals using other techniques

We also focused on the time spent during the individual inspections. Time-related observations were collected by the original experimenters too, but not analyzed. On the contrary, we think that time performance is critical during controlled experiments and can help in interpreting the results.

Although it was not explicitly addressed as a main goal, the original experimenters also analyzed the benefits of inspection meetings. This goal can be stated as

Analyze the inspection process **for the purpose of** evaluation **with respect to** the lost and gained defects after collection meetings **from the point of view of** the researcher.

They discovered that collection meetings produce no net improvement. In other words, the number of defects found for the first time in a collection meeting is equal to the number of defects which are found by the individuals but not reported by the team. We attempt to confirm this result as well in our repolication.

2.2 Experimental Design

To evaluate these hypotheses we replicated the controlled experiment. We ran our experiment in early 1995.

The replication of the experiment manipulates five independent variables:

1. the detection technique: Ad Hoc (AH), Checklist (CH), or Scenario (SC) are used during the inspection;
2. the inspection round: two inspections (R1, R2) are performed by each reviewer;
3. the specification: two SRS (CRUISE, WLMS) are inspected by each reviewer;
4. the order of inspection: first CRUISE (CW) or first WLMS (WC) can be inspected;
5. the team composition: ten 3-people teams perform the inspection tasks.

The detection technique is the treatment variable. The other variables are used to assess potential threats to the internal validity of the experiment. The original experiment included two separate, internal replications. Thus, the original design used the experimental replication as another independent variable. Since our external replication was performed once, we do not include this independent variable.

There are five dependent variables:

1. the team defect detection rate, i.e. the number of defects detected by a team divided by the total number of defects known to be in the specification;
2. the individual defect detection rate, i.e. the number of defects detected by individuals divided by the total number of defects known to be in the specification;
3. the meeting gain rate, i.e. the percentage of defects first identified at a collection meeting;
4. the meeting loss rate, i.e. the percentage of defects first identified by an individual but not included in the report from the collection meeting;
5. the time spent by each participant for the individual inspections.

We collected

Although this last measure was collected by the original experimenters too, it was not defined as a dependent variable and thus it must be considered an additional dependent variable.

Table 1 shows the organization of both the original experiment and our external replication. Teams from the first internal replication are denoted as 1A-1H; teams from the second internal replication are denoted 2A-2H. Teams from our external replication are in bold and denoted A-K. The experimental plan is a partial factorial design in which each team inspects two specifications, one per inspection round, using one detection technique, but not all the combinations of the treatment levels are present.

Detection Technique	**Round 1**		**Round 2**	
	WLMS	**CRUISE**	**WLMS**	**CRUISE**
Ad hoc	1B, 1D, 1G, 1H, 2A **A**, **K**	1A, 1C, 1E, 1F, 2D **D**, **J**	1A **J**	1D, 2B **K**
Checklist	2B **B**	2E, 2G **E**, **G**	1E, 2D, 2G **D**, **G**	1B, 1H **B**
Scenarios	2C, 2F **C**, **F**	2H **H**	1F, 1C, 2E, 2H **E**, **H**	1G, 2A, 2C, 2F **A**, **C**, **F**

Table 1. Experimental plan

In the first inspection round of the first internal replication, all the teams used the Ad Hoc technique. In the second internal replication, the teams that used the Scenario technique in the first inspection round (2C, 2F, 2H) were constrained to use it again in the second round. This was based on the assumption that the use of a systematic technique by a reviewer could affect future performance using nonsystematic

techniques. However, this concern about the possibility of a "carryover effect" can be generalized so that the inspection order conforms to an increasing scale of prescriptiveness. Since the Checklist approach is more systematic than the Ad Hoc approach, we slightly modified the original plan by extending the constraint to Checklist too. As a result, unlike team 2B in the original plan, our team B was not permitted to use the Ad Hoc technique in the second round.

2.3 Participants

The subjects of the original experiment were first and second year graduate students, more than half of them had at least two years of industrial experience. On the other hand, our subjects were third and fourth year undergraduates, just one sixth with at least two years of industrial experience. Our students had experience from a previous software engineering course in SRS reading, but only in the information systems domain. Our students were certainly less experienced than those of the original experiment, but the subjects of the original experiment cannot be considered software professionals either.

We used 10 teams of three subjects for a total of 30 participants, while the original experiment used 8 teams of three subjects for each replication, for a total of 48 participants. The teams in the original experiment were chosen according to different criteria in two different replications. In the first replication, the teams were created with so as to have comparable experience. Based upon a survey regarding background and experience, participants were ranked as low, medium, and high, and then teams were composed by taking a subject from each of the three categories. However, in the second replication team members chosen randomly. In our replication we followed the first strategy. With a low number of participants and a partial factorial design with randomly chosen teams, large differences in ability between teams could occur randomly, thus masking the differences due to detection techniques. However, the survey used as a pretest by the original investigators was not included in the experimental kit, and so we created one ourselves.

2.4 Experimental material

All the material for this replication was present in the experimental kit, although we needed to translate it from English to Italian. Not translating could have introduced a threat to internal validity represented by the difficulty or slowness in understanding the documents.

The material for the experiment included three small SRSs, support documents for the detection techniques, and data collection forms.

Each of the three SRSs described an embedded real-time system: a home temperature control system (HTCS), a water level monitoring system (WLMS), and an automobile cruise control system (CRUISE). The HTCS SRS was used for training in a trial inspection (the original experiment used an elevator control system for training purposes, but the HTCS SRS was included in place of it in the experimental

kit). All the three SRSs adhered to the IEEE format (IEEE, 1984), with an overview written in natural language and the detailed sections specified using the SCR tabular notation (Heninger, 1980). A complete list of the defects for WLMS and CRUISE, but not for the training SRS, was included in the experimental kit. Defects appeared both in the overview and detailed sections of the SRSs.

As a guide for detecting defects, all Ad Hoc reviewers used the same general defect taxonomy. All Checklist reviewers used the same checklist, refined from the taxonomy by adding detailed questions. Scenario reviewers used three different scenarios, derived by refining separate sections from the checklist: data type inconsistencies (DT), incorrect functionalities (IF), and missing or ambiguous functionalities (MF). The experimental kit did not include a classification of the defect list according to the general taxonomy. We classified the defects ourselves but, in order to be sure we were using the same criteria, we also requested the classification directly from the original investigators. Our classification of the defects was identical to the original. Although classifying was a time-consuming activity, this experience can testify to the independence of the defect taxonomy from subjective evaluations. See the appendices in (Porter, 1995) for the taxonomy, the checklist and the three scenarios.

The defect forms, to be filled out by the subjects during the inspections, included various identifiers (SRS, reviewer, and team), the date, the initial and finish time, the kind of activity (detection or collection), the defect location (page and line number), and a textual description of the defect. The defect forms also recorded information the defect disposition (true or false positive). However, only true defects were analyzed in the original experiment and we did the same.

2.5 Preparation and training

We gave five 2-hour lectures on the IEEE standard for SRS, the SCR tabular notation, the inspection process, the defect taxonomy and checklist, and the data collection forms. Lecture references (IEEE, 1984), (Heninger, 1980), (Fagan, 1976) were found in the experimental kit. However, the style for SCR tabular notation in (Heninger, 1980) was partially different from that used in the experimental kit. Thus, we adapted our lectures to the style effectively used for the three SRSs.

After these lectures, we created teams and randomly assigned teams to detection techniques, and subjects to team roles (moderator, recorder, and reader). A trial inspection was performed as a simulation of an experimental session. All the subjects inspected the HTCS SRS using the Ad Hoc technique, as performed in the original experiment and suggested by the guide in the experimental kit. After a two-hour individual inspection and a two-hour collection meeting, another lecture was given as a feedback for the participants.

Afterwards, we gave a lecture on defect-based scenarios but only for those participants who had to use the Scenario technique in the first inspection round. The lecture described the scenarios and showed how they could be applied to the training SRS. After the first inspection round, the lecture on scenarios was

repeated for those subjects who had to use the Scenario technique for the first time in the second inspection round.

2.6 Execution

We conducted the two inspection rounds in a big room with enough space so that the participants would not be disturbed. The students received their experimental package (instructions, SRS, and defect report forms) and had 30 minutes to read and understand the SRS, and to make us questions for clarification. Then, they had 2 hours to do the individual inspection and fill out the defect report forms. After they finished they completed a debriefing questionnaire and returned the material to us.

A short time after all the subjects had finished their individual inspections, the teams met to collect the defects. The collection meeting process is independent of the defect detection technique used for the individual inspection: the reader calls for defects, sequentially scanning the document; the recorder fills out the team defect report form; and the moderator manages the discussion within the team. The collection meetings had a two-hour limit. At the end, the team defect report forms were given back to us.

The two inspection rounds were conducted in a two-day period, while the original experiment used one week. We preferred to shorten the interval period between the two inspection rounds (one day), and the individual-collective inspections (15 minutes) to reduce the threat of internal validity caused by subjects talking to others about their experience.

3. Experimental Results

Looking at the individual and team defect report forms, we validated the defects detected by the participants during the inspections. We considered only true defects, discarding false positives and minor errors, such as stylistic inconsistencies or obviously typographical errors. Our effort was devoted to determining if a defect description, possibly unclear, could be matched to some known defect or could lead to future system faults if remaining uncorrected in the SRS.

In table 2, the set of defects found in the original experiment (Orig) is compared with the set of defects found in our replication (Repl). While almost one half of the original defects were not discovered, our students found new defects not present in the list of the experimental kit. The new defects are shown in table 3 (WLMS) and table 4 (CRUISE). The right columns in these tables show the defect keys used to encode the reviewer's responsibility, i.e. which reviewers we would have expected to find that defect. The AH code indicates an Ad Hoc only defect and thus, excludes the responsibility of Checklist and Scenario reviewers. The CH code means that the defect can be captured using the checklist but not with any scenario, and thus it includes Ad Hoc reviewers (they are responsible for all the defects) and Checklist reviewers. The DT, IF, and MF codes indicates the three scenarios. Each Scenario reviewer is responsible

for its own defect subset; Ad Hoc and Checklist reviewers are also responsible, since a scenario is a refinement of one checklist section.

No. of defects	WLMS	CRUISE
# Orig	42	26
# Repl	34	19
# (Orig ∩ Repl)	25	13
# (Orig ∪ Repl)	51	32

Table 2. Summary of the defects found

Id	Page	Line	Defect	Key
43	3	20	The maximum value that bounds the pump rates is not specified in the glossary.	DT
44	3	22	The Station is not depicted in figure 1 but in figure 2.	CH
45	10	15	With the condition "%watchdog%∈{$uninit$, $shutdown$}", the application variable %control unit% may assume the value $failed$ even if the Controller is not monitoring the Control Unit.	IF
46	13	17	The events "Event(!Test time ≥ 0ms) and Event(!Test time ≥ 2000ms) are not mutually exclusive.	MF
47	14	17	The events "Event(!Test time ≥ 2000ms) and Event(!Test time ≥ 4000ms) are not mutually exclusive.	MF
48	16	17	The events "Event(!Test time ≥ 0ms) and Event(!Test time ≥ 4000ms) are not mutually exclusive.	MF
49	18	8	The input hardware port /clkpulse/ is never used.	AH
50	14	9	The update involves //low window// and not //high window// (the same defect is at page 15, 16 where the update involves, respectively, //level display// and //allarm//).	IF
51	26		Missing reference to figure 3.	AH

Table 3. List of new defects for WLMS SRS

Id	Page	Line	Defect	Key
27	4	14	The gas tank belongs to standard hardware which interacts with the software system.	AH
28	6	4	The system shall relinquish control of the throttle even if the driver presses the accelerator.	AH
29	9	5	It is not specified whether the system knows the mileage.	AH
30	9	22	The event should be "Event(4750 < !oil miles! < 4950)".	MF
31	15	14	The hardware register //average speed display// may assume integer values, but if T >= T0 is true, it assumes the value M/T may not be an integer.	DT
32	16	25	The value is not a non-negative fixed point but a floating-point number in the range [1, 100] as specified on page 25.	DT

Table 4. List of new defects for CRUISE SRS

We did two parallel analyses of the experimental data. The first analysis used only the defects that were present in the experimental kit (Orig). The purpose was to be able to directly compare our results with those of the original experiment. The second analysis used both the defects in the original experiment and

the new defects discovered for the first time by our students (Orig ∪ Repl). The results of both the analyses were roughly the same. Thus, we will show only the results from the second analysis because it provides more data points.

3.1 Team Inspection Performance

The data were analyzed by repeating the analysis strategy of the original experiment. First, data were analyzed with a one-way analysis of variance for each of the five independent variables, to identify the individual variables that could explain a significant amount of variation in the team detection rates. The results, summarized in table 5, revealed a significant effect only for Specification ($p < .05$). Detection Technique, Inspection Round, Inspection Order, and Team Composition did not show significant differences. On the other hand, in the original experiment, Detection Technique was the most significant independent variable ($p < .01$), together with Specification ($p < .01$).

Independent Variable	df	SS	MS	F	p
Detection Technique - main treatment	2	0.00106	0.00053	0.08	0.92
Inspection Round - maturation effect	1	0.00144	0.00144	0.23	0.63
Specification - instrumentation effect	1	0.02450	0.02450	5.03	*0.04*
Inspection Order - presentation effect	1	0.00882	0.00882	1.54	0.23
Team Composition - selection effect	9	0.07235	0.00804	2.02	0.14

Table 5. Analysis of variance for each independent variable

Figure 1 shows the average defect detection rates for both the original experiment and the external replication. Each vertical bar is labeled according to the detection technique and the specification. For example, AH-WLMS represents the average defect detection rate obtained with the Ad Hoc technique when inspecting the WLMS specification.

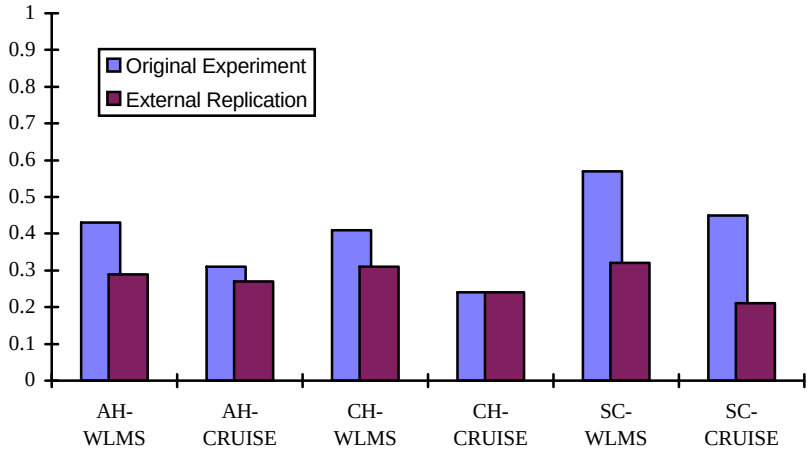

Figure 1. Average team defect detection rates

Figure 2 graphically depicts the variance of the team defect detection rates by showing box-plots for each of the defect detection methods in our external replication. The rectangles include the data points in the first to the third quartiles. The median for each method is identified by a bar within the boxes, while the overall median is represented by the horizontal line across the boxes. Vertical lines outside the boxes extend up from the third quartile to the 95th percentile and down from the first quartile to the 5th percentile. Outliers (not present in our observations) are represented as isolated points. Note that Scenario, the most systematic method, has a higher variance with respect to team defect detection rates.

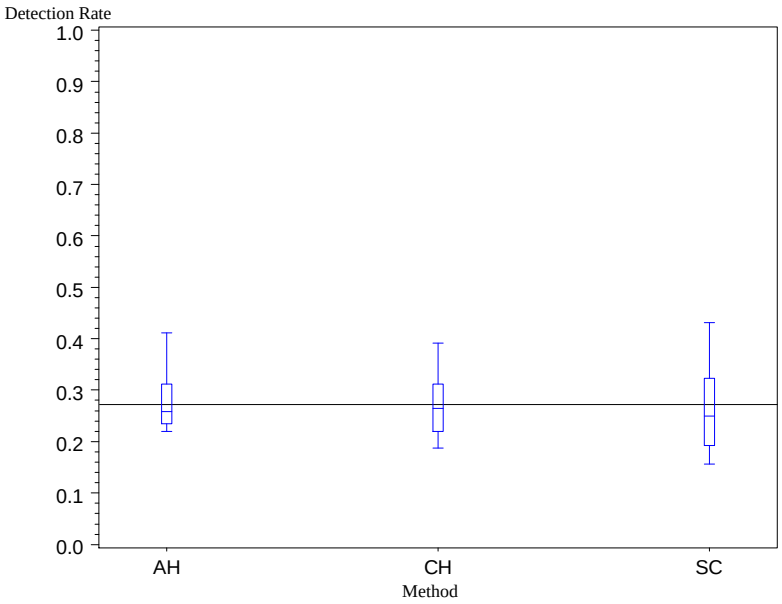

Figure 2. Box plots for the team defect detection rates

As a second step in the analysis of the team performance, we repeated the analysis of the original experiment to see if the main treatment variable, Detection Technique, varied significantly with the different levels of the Specification variable. Data were analyzed using a two-way analysis of variance. As in the original experiment ($p = 0.81$), the interaction between Detection Technique and Specification was not significant ($p = 0.60$).

3.2 Individual Inspection Performance

The analysis of the individual defect report forms helps us to verify if specializing the responsibility of a reviewer results in finding more defects in the related specific class. Furthermore, we also see if the scenario reviewers are equally effective at finding defects for which their scenarios were not designed to detect. If not, reviewer specialization could have an undesired effect when scenarios only have a partial coverage of the defect population, like in this experiment (about 50% of the defects covered by the checklist).

The individual detection rates of Scenario reviewers are compared in tables 6 (WLMS) and 7 (CRUISE) with those of all other reviewers. The right columns in the two tables correspond to the results of one-tail *t* tests performed to verify our hypotheses.

Defect Type	Number Present	Defects found by specific SC reviewers	Defects found by CH reviewers	Defects found by AH reviewers	p-value
DT	15	6.00	4.10	3.60	*0.03**
MF	8	1.50	1.22	0.11	0.10*
IF	7	0.50	0.44	0.33	0.35*
others	21	1.83	4.00	2.55	*0.03***

* null hypothesis: scenario X reviewers do not find any more X defects than non-scenario reviewers
** null hypothesis: scenario reviewers do not find any less 'other' defects than non-scenario reviewers

Table 6. Individual defect detection rates for WLMS SRS

Defect Type	Number Present	Defects found by specific SC reviewers	Defects found by CH reviewers	Defects found by AH reviewers	p-value
DT	12	2.00	2.33	1.77	0.47*
MF	2	0.50	0.22	0.00	NA
IF	3	0.25	0.22	0.11	NA
others	15	1.00	0.88	0.66	0.24**

* null hypothesis: scenario X reviewers do not find any more X defects than non-scenario reviewers
** null hypothesis: scenario reviewers do not find any less 'other' defects than non-scenario reviewers

Table 7. Individual defect detection rates for CRUISE SRS

The results were very different between the WLMS and CRUISE specifications. In the WLMS SRS, the analysis revealed that the number of DT defects found by DT reviewers was significantly higher than the number of DT defects found by non-scenario reviewers ($p < 0.05$). On the other hand, in the CRUISE SRS, the analysis failed to reveal any significant difference.

With respect to the second hypothesis, the analysis of the WLMS SRS revealed that the number of 'other' defects (i.e. defects not covered by any scenario) found by scenario reviewers was significantly lower than the number of 'other' defects found by non-scenario reviewers. On the contrary, with the CRUISE SRS we did not find any significant difference.

For both WLMS and CRUISE, the number of MF and IF defects found by MF and IF reviewers, respectively, was slightly higher than the number of MF and IF defects found by non-scenario reviewers. However, we did not find significant differences for the WLMS SRS, and we could not perform any statistical analysis for the CRUISE SRS, since there were too few data points.

3.3 Collection Meeting Performance

We measured the benefits of collection meetings by comparing the meeting gain rates (i.e. the percentages of defects first identified at a collection meeting) and the meeting loss rates (i.e. the percentage of defects first identified by an individual but not included in the report from the collection meeting). Figures 3 and 4 show the meeting gain and loss rates for WLMS and CRUISE, respectively. Data were analyzed using a *t* test. The analysis failed to reveal a significant difference between the mean meeting gain rate and the mean meeting loss rate ($p = 0.78$ for WLMS and $p = 0.82$ for CRUISE). Thus, our results confirmed those of the original experiment: collection meetings produce no net improvement.

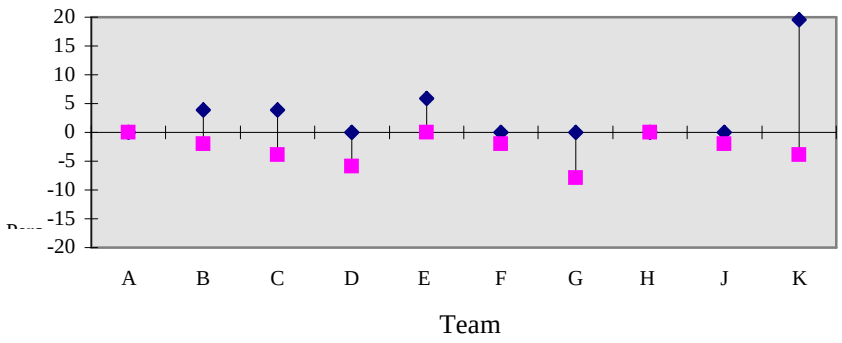

Figure 3. Defect Gain and Loss Rate for WLMS

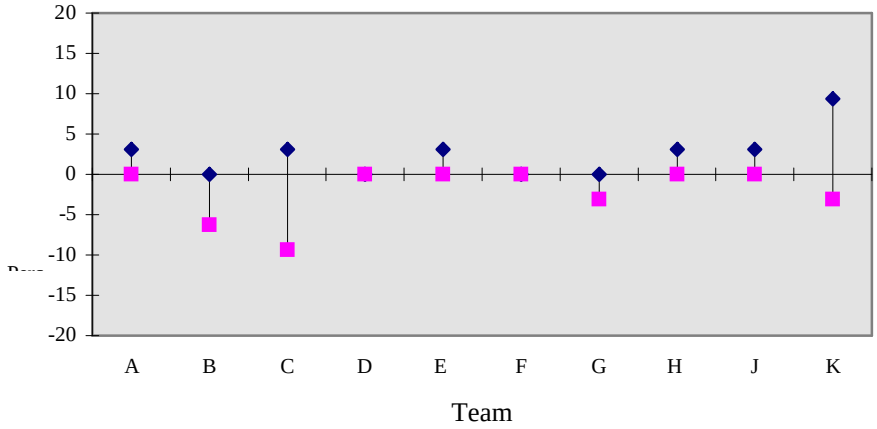

Figure 4. Defect Gain and Loss Rate for CRUISE

3.4 Time Performance

Reviewing the data related to the time spent for individual inspections, we realized that we could not proceed with a statistical analysis of variance to check if there were significant differences between the effects of the independent variables. The time limit given to the subjects was two hours as in the original experiment. Table 8 shows a quartile table of the time spent in individual reviews. The second column provides the percentage of observations which are below the value shown in the third column (in minutes) of the same row. Most of our students finished their individual reviews right at the deadline (more than 50% after exactly 120 minutes and only 25% under 115 minutes). We do not know if the same occurred in the original experiment because time was not included among the dependent variables.

Maximum	100%	120
3rd quartile	75%	120
Median	50%	120
1st quartile	25%	115
Minimum	0%	70

Table 8. Quartiles for time spent in individual reviews

3.5 Questionnaire and Interviews

After the experiment the participants completed a questionnaire built with the purpose of helping us in interpreting the quantitative results. The questionnaire was validated by means of informal interviews performed by a senior student who helped us in running the experiment. Having a colleague as a direct interface enabled the students to feel comfortable in answering our questions. The main results from the qualitative analysis are the following.

- Participants felt they had received too little training in the Scenario approach; some students stated that they really understood how to use a scenario only after the first inspection round. Remember that, as in the original experiment, students performed a training inspection only with the Ad Hoc approach.
- Students complained about the strict time limit; many students declared that they could have done a better job with more time available.
- Participants suffered from an unfamiliar domain, like embedded real-time systems; during their courses, they had been trained mainly on applications in the information systems domain which focuses more on data management and user requests Furthermore, in Italy there are no cars with embedded cruise systems.
- Participants were not able to develop the invariants required by the IF scenario: the state-event model in the training SRS was too simple compared to those in the WLMS and CRUISE SRS.

4. Discussion

With respect to the original experiment, our external replication confirms only that collection meetings produce no net improvement. We cannot confirm the results from the assessment of the defect detection techniques. Controlled experiments work well in establishing a cause-effect relationship when they get positive findings, that is when the null hypothesis is rejected. When there are negative findings, as in this case where no significant differences are found among detection techniques, only more hypotheses can be drawn. In the following we give a list of hypotheses which can explain why the Scenario approach did not increase the overall effectiveness of the inspection process. They can be considered lessons learned that must be carefully considered for future replications of the experiment.

- *Subjects were asked to learn too many new things.* Our students were not familiar with real-time embedded systems neither had they ever driven a car with cruise control. They were not used to modeling control; they learned the SCR tabular notation, and they performed a formal inspection for the first time. From a teacher's point of view, practice with so many issues in a single course can be considered a good result. However, from an experimenter's point of view, uninterpretable results can come from underestimating the subject's learning curve.
- *Defects in the introductory parts create confusion.* Our subjects needed to first understand the problem but they had no error-free introduction. Defects should be contained only in the detailed specifications, leaving the general descriptions free of errors.
- *Training was unfair.* The trial inspection was conducted only with the Ad Hoc approach, while participants could not get confidence with the Checklist and Scenario approaches before the experimental trials. Participants who had the Scenario technique in both inspection rounds said they understood their scenarios only after the first inspection. We cannot verify these affirmations by testing for significant differences in performance, because there are too few observations. However, this can be considered as a threat to the internal validity of the experiment.
- *Subjects who had trouble with the Scenario approach used different techniques to execute the task.* Our students had some difficulties in following the scenarios, especially with the incorrect functionality scenario which required writing invariants from the event and condition tables. A possible interpretation of the greater variance with the Scenario approach could be that some students were successful in applying the scenario, some turned back to Ad Hoc or Checklist, and some continued until the end to apply the assigned scenario but with negative results. The real process should be recorded to prove its conformance to the ideal process.
- *Time limit was too short.* Analyzing human performance in programming activities, Weinberg and Schulman concluded that unreasonably short deadlines would result in erroneous programs, and warned experimenters against mixing the results of subjects who have easily finished with those of subjects who were pressed for time (Weinberg, 1974). We do not know if this happened to the

original experiment. In our case, if we discard all the data from subjects pressed for the deadline there will not be enough observations to analyze.

5. Conclusions

The results of our external replication were different with respect to the first goal of the original experiment, i.e. assessing defect detection techniques, and identical with respect to the second goal, i.e. evaluating the benefits of the collection meetings. The following may be considered the major conclusions derived from our study:

1. The team defect detection rate when using the Scenario technique was not significantly different from those obtained with Ad Hoc or Checklist techniques. The average defect detection rate with Scenario was slightly higher than Ad Hoc and Checklist but only with one specification (WLMS). On the contrary, in the original experiment the defect detection rate when using Scenario was significantly superior (about 35%) to that obtained with Ad Hoc or Checklist techniques. Both the original experiment and our replication found that there were significant differences between specifications (CRUISE more difficult than WLMS) and that there was no significant interaction between detection techniques and specification.
2. During individual inspections of WLMS, the DT scenario was more effective than Ad Hoc and Checklist at finding its specific class of defects. However this is not true for the individual inspection of CRUISE, and we have no evidence for the other two scenarios. On the other hand, in the original experiment all the three scenarios helped to find more targeted defects than Ad Hoc or Checklist. Furthermore, for the WLMS specification, the Scenario technique showed a lower ability in finding non-targeted defects than Ad Hoc and Checklist, while for the CRUISE there were no significant differences. In the original experiment, the Scenario technique was as effective in finding other classes of defects as Ad Hoc and Checklist.
3. Both in the original experiment and in our external replication, collection meetings did not provide a net meeting improvement because meeting gains were offset by true defects lost in the meeting.

We discovered problems, some related to the change in experimental environment but others related to the experiment itself, that could be considered threats to internal validity. These include learning curve, unfair training, task conformance, and time limit. Although we are not able to explain exactly why we got partially different results, we can offer some reasonable hypotheses and suggest improvements to enable other researchers to attempt new external replications.

Acknowledgments

We gratefully acknowledge the collaboration of Pier Francesco Fusaro in the execution and the analysis of the experiment. Our thanks to all the students participating in the experiment for their hard work. We

would like to thank Vic Basili, Adam Porter and Larry Votta for their useful technical comments, and all ISERN (International Software Engineering Research Network) members for having encouraged this replication. Thanks also to Carolyne Seaman, Forest Shull, and Sivert Sorumgard for having improved a draft version of this paper.

This work has been partially supported by the 40% funds of the Italian M.U.R.S.T. under the project "V&V in software engineering".

References

Basili, V. R., Caldiera, G., and Rombach, H. D., 1994. Goal Question Metric paradigm. *Encyclopedia of Software Engineering*. Marciniak J. J. (ed.), Ney York, NY: John Wiley & Sons.

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

Heninger, K. L., 1980. Specifying software requirements for complex systems: new techniques and their application. *IEEE Trans. Soft. Eng.*, SE-6: 1, 2-13.

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

IEEE Std.830-1984. *IEEE Guide to Software Requirements Specification*. Soft. Eng. Tech. Comm. of the IEEE Computer Society.

Judd, C. M., Smith, E. R., and Kidder, L. H., 1991. *Research Methods in Social Relations*, 6th edition. Orlando: Holt Rinehart and Winston, Inc.

Parnas, D. L., and Weiss, D. M., 1985. Active design reviews: principles and practices. *Proc. 8th Int. Conf. Soft. Eng.*, 215-222.

Porter, A. A., Votta, L. G., and Basili, V. R., 1995. Comparing detection methods for software requirements inspections: a replicated experiment. *IEEE Trans. Soft. Eng.*, 21:6, 563-575.

Weinberg, G. M., and Schulman, E. L., 1974. Goals and performance in computer programming. *Human Factors*, 16:1, 70-77.