Improving the Mechanisms for Replicating Software Engineering Experiments

Natalia Juristo, Sira Vegas, Ana M. Moreno
Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo,
28660 Boadilla del Monte, Madrid, SPAIN
+34 91 336 6922, +34 91 336 6929
{natalia,svegas,ammoreno}@fi.upm.es

Patricio Letelier
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, SPAIN
+34 996 387 7007 ext. 73589
letelier@dsic.upv.es

ABSTRACT

As in other spheres of science and technology, the replication of experiments in SE is an indispensable task. However, replication is extremely difficult in SE. This is primarily due to the complexity of the context in which experiments are run. The success in accurately describing the context of an experiment will be later reflected in the possibility of combining the results of the replications. If the context is not properly characterised, it will be impossible to isolate the variables causing any discrepancies between the results of the different replications. In this paper, we have used and then compared the instruments for transmitting information about experiments proposed in the literature to conduct replications of SE experiments. Based on this experience, we propose an improved instrument.

Categories and Subject Descriptors

D.2.0 **[Software Engineering]:** General

General Terms

Experimentation

Keywords

Empirical Studies, Experimentation, Replication, Combination of experiment results.

1. INTRODUCTION

Experiment replication is a key feature of experimentation in any scientific or technological field. Replication involves *other researchers in other settings with different samples attempting to reproduce the research as closely as possible* [6].

Although this definition reflects the concept of replication in other sciences, it is difficult to attain close (ideally exact) reproductions in experimentally immature fields like SE.

Despite this difficulty, experiments need to be replicated to configure an experimentally backed body of knowledge. This body of knowledge is generated by integrating the results achieved in different replications of experiments. For this purpose, researchers first need to look at whether the results of these replications are (totally or partially) consistent. Second, they need to analyse the reasons for any such convergences or divergences to gain an insight into and find out what variables cause them. Third, they need to generate pieces of knowledge specifying the circumstances under which they are applicable.

A number of attempts have been made at combining experimental results in SE ([5], [8], [13], [14], [15], [17]). The fruits of all the attempts, both trials employing statistical techniques and studies using more informal approaches, have been very disappointing. In the case of the informal combinations ([5], [8], [17]), the findings have been very limited, because of the discrepancies among the results of the replications and the impossibility of isolating the reasons for the discrepancies due to context variations among replications. The attempts at combination using statistical techniques ([13], [14], [15]) have turned out to be impracticable. Only when the same researchers have attempted to statistically combine the results of their own replications ([12]) have they been successful. This confirms that when the context of an experiment is reproduced accurately (which is much easier with the same researchers in the same settings), combination of experimental results appears to be feasible.

The source of the main obstacles to combining experimental results is the intrinsic difficulties of replications in SE. Exactly reproducing experiments in SE is very difficult, if not impossible, since experimental conditions are seldom identical. It is very unlikely that researchers will have access to the same resources (number of subjects, time, etc.) or be able to track down subjects that are equally knowledgeable about the technologies that are being experimented with, etc.

Therefore, one of the major difficulties facing SE researchers when replicating experiments is to reproduce experimental conditions in different settings. The context of a SE experiment is very complex due to the very many variables involved in the phenomenon under examination. It is practically impossible to control all variables in a SE replication. Because the context is complex, a lot of information about the original experiment is needed if it is to be satisfactorily replicated. For example, when replicating experiments on different SE techniques, researchers need to know not only which techniques were examined, but also how the techniques were applied, how the subjects were trained, what knowledge these subjects already had, etc.

A number of instruments for transmitting information have been proposed to improve the running of replications in SE. Their aim is to describe the experiments in as much detail as possible to allow an accurate reproduction of the experiment. This article reflects our experiences using some of these instruments to

Conference'04, Month 1–2, 2004, City, State, Country.

perform replications. As a result of the lessons learned from this experience, we have been able to propose an information transmission instrument that improves both the replication and later aggregation of results. We used our proposal in another replication, whose results are also reported here. We made use of these results to fine tune the original proposal.

The article is organised as follows. Section 2 presents the instruments for transmitting information proposed in the literature. Section 3 describes the history of the experiment that we have used as a benchmark. As part of our research, we have run a further two replications with the aim of examining the information transmission instruments. These two replications are described in sections 4 and 5, respectively. These sections also analyse the problems derived from the instruments used and propose some solutions to these problems. Section 6 summarises our proposal for an information transmission instrument for the purpose of performing experiment replications in SE. Finally, section 7 presents the conclusions of this research.

2. INSTRUMENTS FOR TRANSMITTING INFORMATION FOR REPLICATION

In SE, different manners of transferring information among researchers have been proposed for the purposes of replicating experiments. These are what we term **Information Transmission Instruments (ITI)**, which we divide into:

- *Documentation necessary for running the replication.* This is what usually goes by the name of replication package, experimental package or laboratory package. We use the term **Replication Package (RP***).*
- Setting up of *communication channels* among groups of earlier researchers and replicating researchers. We use the term **Inter-Researchers Communication Mechanism (ICM**) to refer to how this communication is effected.

On this basis, the ITIs for replication can be classed according to two parameters: RP contents and type of ICM.

There are a number of proposals regarding what information a RP should contain. Nevertheless, the question cannot yet be considered to be settled, since new improved proposals are being put forward all the time. The idea of what documentation should be transferred to the replicating researchers has changed over the years, as have its contents[1]:

- In the early days, the documentation available about an experiment consisted of **articles** about the experiment. There was no associated RP. This is the information used as a basis for the replication in [11], for example.
- When the concept of RP was introduced, this was defined as the material required during the operation of the experiment (documentation to be delivered to subjects, data collection mechanisms, etc). Note that the definition, planning and operation of the experiment, as well as the analysis of the collected data are not reflected in this type of RP but in articles about the experiment. We term this type of package **Operational RP**. This RP type was used, for example, in [16].
- This conception was later extended, as proposed in [4], and later used, for example in [18], to include the definition and planning of the experiment, the data collected during the experiment and the material required for training, apart from the contents of the Operational RP. Note, however, that the analysis of the data was still reflected in articles about the experiment. We term this type of package **Descriptive RP**. According to this proposal, an experiment is not as an isolated event, but a part of a family[2]. Therefore, the RP also contains the aggregation of the results of earlier replications.
- In [18], a RP is proposed which, apart from contents of the Descriptive RP, includes the procedures associated with running the experiment (for example, guidance specifying what tasks the experimenter should perform during the operation of the experiment), as well as a distinction between specific parts associated with the replications and generic parts associated with the experiment to provide for the growth of replications. We term this type of documentation **Procedural RP**. We have found no published replications that make use of this type of package to date.

As regards the ICM, earlier experimenters have communicated with replicating researchers in several manners[3]:

- The simplest mechanism is no communication among researchers. This mechanism is used, for example, in [16]. In these cases, the transmission of information among the researchers is confined to the transfer of documentation. That is, there is no other type of interaction aside from the mere transfer of the RP. We term this manner of interacting **Zero ICM**.
- Another slightly improved mechanism is used in [11], in which earlier researchers settle occasional doubts for the replicating researchers. We term this manner of interacting **ICM with Query Answering.**
- Sometimes, there is occasional collaboration among researchers, such as, for example, the replicating researchers visiting the earlier researchers while they are performing a replication, or earlier researchers analysing the data collected by the replicating researchers. This mechanism is used in [3], [17]. We term this manner of interacting **ICM with Occasional Collaboration.**
- Finally, the most complex communication mechanism involves collaborative work among researchers, as proposed in [18]. Here, the authors describe several replications run by different groups of researchers. The ICM in this case is composed of different types of workshops (virtual and presential), e-mail, web portals and a knowledge repository. The cooperation takes place among all the groups of researchers. Considering the close collaboration between researchers that is proposed, we term this manner of interacting **ICM with Heavy Collaboration**.

In this paper, we propose an improved ITI. Our proposal is composed of:

- A **Family RP**. This RP improves the Procedural RP in several ways: it includes some new items, it refines some other items and, above all, it contains a new structure, especially designed to ease the aggregation of the results of different replications.

[1] In their articles, experimenters do not always explicitly specify what RPs they used. In these cases, this has been inferred from what the researchers state in the articles reporting the replications.

[2] Family means a set of replications of an experiment.

[3] In their articles, experimenters do not always explicitly specify what ICMs they used. In these cases, this has been inferred based on what the researchers state in the articles reporting the replications and/or whether or not the researchers participating in different replications have co-authored articles.

- An **ICM with Agile Collaboration**. This an ICM halfway between the *ICM with Occasional Collaboration* and the *ICM with Heavy Collaboration*, since the *Occasional Collaboration* is insufficient and the *Heavy Collaboration* is too taxing. We consider the *Heavy Collaboration* to be too demanding for several reasons. First, it requires close collaboration among all the researchers that run the replications for a long period of time (several years in the particular case of [26]). Second, the context of the replications must be exactly the same as in the original experimental setting. Finally (in the case of [26]). the different groups of researchers are known to each other as they have worked together before. From our experience, these requirements are sometimes difficult to meet. Our proposal aims to: make it easier to tailor the replication to the new context with the aid of earlier researchers so that any changes have the least possible impact on the results; require the minimum amount of collaboration possible; and foster the collaboration among researchers that have not worked together.

We have obtained our IT proposal by performing replications of an experiment whose history is presented in section 3. For the first replication, described in section 4, we employed the *Operational RP* that the last replicating researchers had prepared and *Zero ICM*. From this experience, we were able to come up with a preliminary proposal for a *Family RP* and *ICM with Agile Collaboration*, which was tried out during the second replication, described in section 5.

3. DESCRIPTION OF EARLIER REPLICATIONS

To understand what our replications involve, earlier replications need to be described. For reasons of space, this description is founded on the results of these replications and the ITIs used. The full details of these replications are given in [19].

The goal of these experiments is to examine the relative effectiveness of different code evaluation techniques. Basili and Selby [1], [2] ran the original experiment, plus two replications. Then Kamsties and Lott [10], [11] ran another two replications. Later, Roper, Wood, Brooks and Miller [16], [20] ran another replication.

The goal of Basili and Selby's replications [1], [2] was to analyse the effectiveness, cost, and number of faults detected by fault type. For this purpose, three factors were defined in the experiment: code evaluation techniques, software type, and three expertise levels. The SIMPL-T programming language was used in the first two replications, whereas FORTRAN was used in the third. Basili and Selby's replications used a fractioned factorial design, in which each subject applied the three techniques to three different programs. In all cases, subjects were set the task of first applying the technique and then, for boundary analysis and statement coverage, running the generated test cases.

Kamsties and Lott [10], [11] describe their experiment (referred to hereinafter as the KL replication) as a replication and extension of Basili and Selby's experiments, but, in actual fact, the only things they do not alter are the goal and the basic design. More specifically, the changes they made were as follows. They examine none of the response variables except defects detected by subject, whereas they add faults isolated and time per stage (test case generation, failure identification and fault isolation). They switch the sentence coverage technique for predicate coverage, as well as the programs used. This time the subjects have exclusively a junior expertise level. They introduce a new factor, called group, to examine what influence the order in which the subjects apply the techniques has. In this case, the programs are in C,. We assume that the ITI used for this replication was *publications*[4], plus *ICM with Query Answering*[5].

Roper *et al.* [16], [20] run a replication of Kamsties and Lott's experiment (which will be referred to hereinafter as the RWM replication). Their changes were: the use of defects detected as the only response variable, switching the technique under study from predicate coverage to branch coverage, and not using the group factor. The ITIs between Kamsties and Lott and Roper *et al.* are: *Operational RP*[6] and *Zero ICM*[7].

We have tried to combine the results from the replications described here, the outcome of which can be found in [19]. However, it was fruitless. We think that this was because the ITIs used to carry out the different replications prevented the original experiment from being accurately reproduced. To confirm this hypothesis, we ran another replication, using the instrument of the last replication —*Operational RP* and *Zero ICM*—, which is described in the following section.

4. REPLICATIONS WITH OPERATIONAL RP AND ZERO ICM

This section shows our experience in running further replications of the earlier experiments. Specifically, the replications were conducted by Juristo and Vegas [9] (and are referred to hereinafter as the JV replications). In the following, we describe how the replication was run in terms of the ITIs used, as well as the problems encountered.

4.1 Information Transmission Instruments Used

To run these replications, the ITI used was the available *Operational RP* and *Zero ICM*. The contents of the RP were[8]:

- Specification of the training and experiment programs.
- Source code with training and experiment program faults.
- Description of the faults in the training and experiment programs, together with the failures caused when run.
- Solution for each of the three training programs for different techniques.
- Instruction sheets for subjects per program and technique.
- Data collection forms.

Because we were using an *Operational RP*, we also had to consult

[4] The original experiment is detailed in [1] and [2], and no other material has been found to have been put together for replication purposes.

[5] The only sign that there is communication between researchers is Kamsties and Lott's acknowledgement of the original authors in [16] for answering some queries.

[6] Experimental package that Kamsties and Lott put together [10].

[7] The only sign of any communication between researchers is Roper *et al.'s* acknowledgement of Kamsties and Lott in [16], [20] for the RP.

[8] We used the RP developed by Roper et al., whose contents were the same as Kamsties and Lott's RP, tailored to their replication. To access the RP, we contacted Roper, who provided both Kamsties and Lott's and their own package.

publications [10], [11], [16], [20] to understand the pre-[9] and post-[10]experimental stages, as they contained information not included in the RP.

4.2 Running the Replications

Although the JV replications were designed as an exact replication of the RWM replication, a series of changes were made due to variations in the context of the experiment. These changes were:

- The functional technique used was changed. Equivalent class partitioning was used in place of boundary value analysis[11].
- The fault isolation step was eliminated. In the RWM replication, the authors stated that they were unable to analyse the data concerning this step because hardly any of the subjects had time enough to complete the task. As our subjects did not have much more time to do the experiment than the subjects in the RWM replication, we opted to remove this step.
- The follow-up phase was also removed, because time was short.
- Likewise, before performing the replication, the package material to be used by the experimental subjects (programs, forms, etc.) was translated to Spanish to rule out any bias being introduced into the experiment.

Therefore, the changes made were due mainly to differences in our setting with respect to previous replications: resource availability (time) and subjects characteristics (language).

There can be no doubt that the use of a RP is vital for being able to run a replication. Without it, it would not have been possible to use the same programs with the same faults, the same data collection forms and the same training exercises. It is essential that all these conditions are reproduced to be able to speak of experiment replication.

However, the replication of SE experiments is something that is so complex that the use of an Operational RP does not appear to be sufficient. When we performed the replications, we came up against a number of problems that are summarised in Table 1. This table also shows the solutions for these problems, and whether they have already been proposed:

- We would have liked to have discussed the decision to change the technique with earlier experimenters. To solve this problem, we propose communication between researchers, conducting a joint study of the changes to be made to the definition and planning of the experiment owing to variations in the context of the replication setting.
- The package did not include information on the definition and planning of the experiment. Because it was reflected in a publication, its length was limited. It would be impossible to get this information in the case of replications of unpublished experiments. The Descriptive RP would have solved this problem, since it includes the definition and planning of the experiment.
- Although we had the training programs used in the earlier replications, and the earlier experimenters gave a brief description of the techniques used in the experiments, we did not know what teaching material was used to train the subjects. The Descriptive RP would have solved this problem, since it includes the training material.
- There was no guidance as to the sequence of the tasks to be performed during the operation of the experiment. It would have been helpful to have a more detailed description of the operation of the experiment, especially of the procedure to be followed when running the experiment. The Procedural RP would have solved this problem, since it includes a detailed script of what the experimenter should and should not do when running the experiment.
- We would have appreciated a more detailed description of what the programs do, as some points of some of the specifications turned out to be ambiguous. To solve this problem, the program specifications should be improved.
- We had doubts about how some of the data collection forms were to be filled in. We propose including examples of how to fill in the data collection forms.
- We would have liked to have the source code of the programs without faults, as it was not always so evident how they were to be corrected. The Procedural RP would have solved this problem, since it includes the faultless source code.
- We would have liked the experimental material to have been organised in files so as to minimise the material preparation time for delivery to subjects and for error prevention. We propose improving the organisation of files that contain the data collection forms to ease the preparation for running the experiment.
- We would have liked the RP to have been self-contained and not have had to look for information about the operation of the experiment in publications about the experiment. The Descriptive RP would have solved this problem, since it includes the description of the experimental operation.

4.3 Interpretation and Aggregation of Results

The goal of replication is to mature the experimental body of knowledge. Therefore, it is essential to integrate the results of the replications. When we tried to aggregate the results of the JV replications with the KL and RWM replications, we came up against the following problems:

As regards **detected defects**, the results were:

- The **equivalence class partitioning technique behaves identically to branch coverage and both perform better than code review**. This result *is not coherent* with the results of the earlier replications and could not be aggregated.

  In the KL replication, it was found that the three techniques behaved equally. On the other hand, the RWM replication claims that there is a dependency, but this was not explored.

  We suspect that the reason for this divergence in the results is to be found in the fact that the training the subjects received in the techniques was different in all replications.
- The **equivalence class partitioning and branch coverage techniques are highly effective**. This result is *not coherent* with the results of earlier replications and could not be aggregated.

  The two dynamic techniques behaved better than the respective functional and structural techniques in the KL and RWM replications. Subjects appear to be applying the code review technique worse and the other two techniques better than in the previous replications.

[9] Experiment definition and planning.

[10] Analysis of collected data.

[11] We thought that it was the most commonly used technique in professional practice.

Table 1. Problems found during JV replications, along with their solutions.

EXPERIMENT STAGE	PROBLEM	SOLUTION	NOVELTY OF THE SOLUTION
RUNNING THE REPLICATION	No chance to discuss certain changes with earlier researchers	More communication	Joint study of new context
	Definition and planning not included	New item in RP	Descriptive RP
	Teaching material not included	New item in RP PR	Descriptive RP
	Tasks to be performed by the experimenter during the operation not known	Improve item in RP	Procedural RP
	Insufficiently detailed description of program specifications	Improve item in RP	Include more detailed specs
	Instructions for filling in data collection forms not included	Improve item in RP	Include examples
	Correct (without faults) source code not included	Improve item in RP	Procedural RP
	Better way to organise experimental material files	Improve item in RP	Operation material files
	Description of operation of experiment not included	New item in RP	Descriptive RP
INTERPRETATION AND AGGREGATION	Impossibility of identifying the sources of variability	More communication	Joint aggregation
	Some analyses not included	New item in RP	Descriptive RP
	Data analyses of earlier experiments not included	New item in RP	Include specific data analysis
	Combination results with earlier experiments not included	New item in RP	Descriptive RP

A possible reason for this divergence could be the subjects' previous knowledge. Code review is perhaps a technique with which computing students are less familiar than testing techniques. Students learn the philosophy behind testing informally in programming courses. This may lead to the techniques not competing on equal terms (some informal knowledge of the technique vs. completely unfamiliar technique).

- **The programs do not behave equally**. This result is *not coherent* with the results of earlier replications and could not be aggregated.

  The JV replications and the first KL replication find that there is a dependency on the program, although with contradictory results. In the RWM replication, a program dependency is also found, although it was not explored. Finally, in the second KL replication, no program dependency is observed.

  In this case, we are unable to identify what variable might be causing the discrepancy in the results. It could perhaps be due to experimental error, subjects, their training or to some uncontrolled change having occurred during the experiment operation.
- **There is an interaction between program and technique.** This result is *not coherent* with the results of earlier replications and cannot be combined.

  This interaction is observed in the JV and RWM replications, but not in the KL replications. The RWM replication does not identify the interaction type, as it is not examined by the experimenters. In the JV replications, it is found that, although the difference in behaviour between the dynamic techniques and the static technique remains constant, the techniques may behave slightly better or worse depending on the program.

  Again we are unable to identify what variable might be causing the discrepancy in the results. As above, any cause is plausible.
- **There is a dependency on the faults found by type**. This result is *not coherent* with the results of earlier replications and cannot be combined.

  In the JV replications, it is found that this dependency only occurs for the cosmetic type faults, whereas one of the KL replications claims that it does not depend on fault type, and, according to the RWM replication, the difference in behaviour is confined to the dynamic techniques, which behave better than the static technique for faults of omission and control. The RWM replication concludes that there is a dependency with respect to the fault type, but their analysis does not explain what this dependency is.

  Again we are unable to identify what variable might be causing the discrepancy in the results.

As regards the **efficiency** aspect[12], it was found that:

- **The functional technique is less time-consuming than the structural technique, and this is less time-consuming than code review.** This result *is coherent* with the results of KL replications and could be aggregated.
- **The time ratio is 1.4 for the functional and structural techniques and 2 for the functional and code review techniques.** This result could not be aggregated with the results of earlier replications, because, in the analysis of results of the KL replications, the technique application time (although it was measured) was reported in conjunction with the fault detection time rather than separately.

The problems encountered with the replication in this stage have been (see Table 1):

- It has been impossible to isolate sources of variability to explain the divergences between the results of the different replications. This was because of uncontrolled sources of variability, which have led to multiple possible causes of the divergences when trying to interpret the results of the replications. We propose that the two groups of researchers should combine the results of the different replications jointly.
- Some analyses of the results of earlier replications were missing in the publications (for instance, technique efficiency). The Descriptive RP would have solved this problem, since it includes the data collected during the experiment. In this manner, if the replicating researchers find any type of analysis to be missing, they can do it afterwards.
- Again it would have been helpful for the RP to have been self-contained and have included full information about the data analysis. We propose that the results of the data analysis be included.
- It would have been helpful for the RP to include information related to the aggregation of the results of the earlier replications. The Descriptive RP would have solved this problem, since it includes the results of the combination with earlier replications.

[12] Note that this aspect was not examined by Roper *et al.*

The JV replications were run a total of five times at Madrid Technical University (UPM) from 2000 to 2004. The results of these replications run in the same setting with different samples by the same experimenters were always coherent and easily integrated. Like [12], we confirm that replications run by the same experimenters can be aggregated. Therefore, aggregating replication results in SE is possible, provided the sources of context variation of the different settings are controlled.

4.4 Evaluation of the Information Transmission Instrument

Using an *Operational RP with Zero ICM* experimenters are thoroughly acquainted with the experiment (as experimental material is available) and can aggregate results in those cases where there is no divergence. However, as is clear from Table 1, we encountered problems throughout the running of the experiment, starting with the definition of the new replication and ending with the aggregation of the results.

The most striking finding from the attempt at integrating the results of the JV replications with the KL and RWM replications is that, in some cases, aggregation was impossible. It is worth noting that we are talking about replications of the same experiment where a RP has been transferred between researchers. Even so, we were unable to identify the source of variability that could have caused the discrepancy in the results. We are convinced that the reason was the problems encountered when running the replication. Therefore, this ITI does not allow exact reproduction of an experiment.

We have set out to improve the ITI used during the JV replications for future replications. From Table 1 we can see that there are two possible ways of improving the Operational RP: introduction of new items in the RP and improvement of the description of an item present in the RP. Around 50% of the improvements we proposed to the Operational RP type have already been accounted for by the Descriptive and Procedural RPs. This is not surprising, as these RPs were proposed after the RP used in this replication had been prepared.

However, it seems to us that even the most extensive RP is not sufficient when the new context is not the same as in the earlier experiment (which will most often be the case), because an RP is of no use either for adapting the experiment to a new context or for identifying the sources of variability during the aggregation. Zero ICM does not allow exact replications. Much more communication is necessary. We looked at Occasional and Heavy Collaboration as possible ICMs for use in a new replication, this time in a different setting. As mentioned in section 2, Occasional Collaboration seemed to us to be too weak an ICM for solving the problems that we had encountered in the JV replications (listed in Table 1). On the other hand, Heavy Collaboration was more demanding than what either we or the replicating researchers could undertake. So, we developed an ICM, which we termed *ICM with Agile Collaboration*, that is midway between these two in terms of the demands it places on researchers.

5. REPLICATION WITH FAMILY RP AND ICM WITH AGILE COLLABORATION

Another replication of the experiment was performed in 2005. This replication was to be run by different experimenters, namely researchers from the Technical University of Valencia. We refer to this replication hereinafter as UPV replication[13].

The way in which the tasks were divided among researchers in this replication was: the definition and planning of the experiment was done jointly by the UPM (earlier researchers) and the UPV (replicating researchers), the UPV undertook the operation of the experiment, and the UPM analysed the data. The results were combined with earlier replications jointly by the UPM and UPV. Finally, the UPM put together the RP.

5.1 Information Transmission Instruments Used

For this new replication, we developed a *Family RP* that included the improvements proposed in Table 1 Likewise, an *ICM with Agile Collaboration* was set up, following the proposed solutions listed in Table 1, as none of the proposals discussed in section 2 was suitable under the circumstances.

As regards the RP put together by UPM, it was composed of two general parts that described the experiment and aggregated the results of the different replications respectively, and two specific parts, one of which described the JV replications and the other the UPV replication. The contents of the RWM package, modified for the contexts of the JV and UPV replications, were used to put together the specific parts. Each specific part included the contents suggested by the *Family RP* and was composed of:

- The definition and planning of the replication.
- Detailed description of the operation of the experiment, including a script with the description of all the tasks to be performed by the experimenters.
- Material needed for the operation of the experiment. This includes:
  - – Training material (slides and bibliographical references).
  - – Data collection forms.
  - – Instructions sheets for subjects per program and per technique.
  - – Specification of the training and experiment programs.
  - – Training and experiment program source code with and without faults.
  - – Solution for each training program.
  - – Examples of how to fill in the data collection forms.
- Description of the faults in the training and experiment programs, alongside the failures they cause when run.
- Electronic material needed for operation of the experiment. This time the material was organised by sessions to ease the preparation of the experiment operation.

Additionally, the ICM with Agile Collaboration consists of:

- A meeting for the definition and planning stage. The purpose of this meeting is to analyse the context in which the new replication is to take place.
- Provision is made for the possibility of making queries over the telephone or by e-mail to settle occasional doubts during experiment operation.
- A second meeting is held for results aggregation to look jointly for the causes of divergences between the JV and UPV

[13] We opted to run a replication at a Spanish university, because the similarity of some aspects of the setting allowed us to control some variables concerning the subjects, like previous training or native language.

replications.

5.2 Running the Replication

During the definition and planning meeting, it was found that the replicating researchers did not have the time it would take to run the experiment as in JV replications. This meant that changes had to be made to the definition of the experiment operation to tailor the experiment to the new context, where less time was available. The changes that were made for this replication involved:

- Eliminating the code review technique.
- Each session worked with one technique and three programs (rather than one program and all techniques)
- Test case generation and test case running parts were separated into different sessions. The subjects ran test cases for one of the programs rather than for the two on which dynamic techniques were run.
- The training of the subjects was altered. In the JV replications, the subjects were given lectures on the techniques, because they were unfamiliar with them. In the UPV replication, the subjects were already acquainted with the techniques that were taught in another course. Then, in the UPV replication, training was confined to a refresher tutorial on the techniques in the shape of a practical exercise.

The decisions concerning the changes to be made in the experiment were taken jointly by researchers from the UPM and the UPV. A joint decision supervised by researchers that have already run the experiment several times should improve later results aggregation, preventing what happened in the aggregation of the JV with the KL and RWM replications.

In the following, we describe the problems encountered during the running of the replication. Table 2 gives a summary of these problems, as well as a proposed solution for each one. It should be noted that these problems were discovered at the results aggregation meeting. However, as the problems are related to this stage, they are described here:

- The replicating researchers did not use the training material included in the RP at all. The material that was provided in the package was not tailored to the UPV setting. All the training was delivered, when just the training programs would have been sufficient, as the students were acquainted with these techniques. We propose tailoring training to the new replications context.
- The replicating researchers found the organisation of the electronic data collection forms to be complicated. It should be noted that the UPM put together a session-based organisation. However, the replicating researchers specified that they found this organisation to be too coarsely grained. They would have preferred an organisation itemised by experimental design group within each session. We propose reorganising the files that contain the experimental material tailored to the replication setting.
- The replicating researchers would have liked to have had a description of the atmosphere of the experiment. They had doubts about how they were to treat the subjects during the experiment. For example, were they to answer questions about the techniques, programming language or programs? Were subjects to be allowed to talk or use their notes while the experiment was being run? This indicates that the script provided in the RP for the replicating researchers was not detailed enough. So, we propose improving the script for the replicating researchers by introducing information about what the experiment atmosphere should be like**.**
- The subjects had doubts about how to fill in the data collection forms. Although the RP included examples of how to fill in the forms, it was not stressed that this was not only for the experimenters but should also be explained to the subjects. Again, this indicates that the script for the replicating researchers provided in the RP was deficient. So, we propose improving the script for the replicating researchers by introducing what things experimenters should and should not do.
- There was a scheduling error for the application of the structural technique and, as a result, the structural subjects did not have time to complete the application of the technique. We propose holding a second meeting after sending the RP to the replicating researchers and before running the experiment for the purpose of giving explanations and better coordinating and sharing the researchers' view of the experiment operation. Apart from this point, some of the other problems that have been detected might have come to light at such a meeting.

5.3 Interpretation and Aggregation of Results

To aggregate the experimental results we held a meeting between both groups of researchers. The goal of this meeting was to jointly interpret the results of the JV replications and the UPV replication. Some divergences were found during the analysis of the UPV replication data and the later attempt at combination with the results of the JV replications. We wanted to discuss the possible sources of these divergences. The importance of this meeting lies in the discussion of the integration of the results of the replications, by researchers that have been present during the operation of each replication and are well acquainted with the context and settings of each replication.

As regards the **relation between the technique and fault type**, it was found that:

- The **effectiveness of the equivalence partitioning and branch coverage techniques is similar, irrespective of fault type**. This result *is coherent* with the results of the JV replications and could be aggregated.

Table 2. Problems found during the UPV replication, along with their solutions.

EXPERIMENT STAGE	PROBLEM	SOLUTION	NOVELTY OF THE SOLUTION
RUNNING THE REPLICATION	Training material was not tailored to replication	Improve item in RP	Tailor training material
	Complexity of the preparation of the operation material	Improve item in RP	Improve organisation of material
	The atmosphere of the experiment was not included	Improve item in RP	Include atmosphere of experiment
	There was a scheduling error	More communication	Pre-execution meeting
	Details were missing from the experiment script	Improve item in RP	Add to experiment script
INTERPRETATION AND AGGREGATION	None	--	--

In both the JV and the UPV replications, the percentage of defects that a subject is capable of detecting matches for the two techniques under study. This result ties in with the findings of the KL replication, albeit not the RWM replication.

- The **effectiveness of both techniques is medium**, lower than in the JV experiment. This result *is not coherent* with the results of the JV replications and could not be aggregated.

  The percentage of defects that a subject is capable of detecting falls considerably with respect to the JV replications. This result matches the results of the KL and RWM replications.

  Initially, we had no explanation for this discrepancy with respect to the JV replications. We suspected that it might be due to training, but this did not appear to be logical, as the RP contained the material necessary for training the subjects during the experiment.

  The meeting with the UPV researchers helped to confirm that the cause was training. In the UPV replication, subjects were simply reminded of how to use the techniques by means of a practical example, which was simpler than the ones given in the RP, as they had taken a course on the techniques in an earlier year and were supposed to know how to use them. However, the UPM subjects had received several training lectures just before the experiment and the details of the techniques were fresher than for the UPV subjects.

As regards the **efficiency** of the techniques, it was found that:

- The **time it takes to apply both techniques is different**. This result *is coherent* with the results of the JV replications and could be aggregated.

  The functional technique does not take as long to apply as the structural technique. It should be noted that the RWM replication does not examine this point, and the KL replication did not provide the results of the analysis.

- The **time it takes to apply the functional technique is abnormally low**. This result *is not coherent* with the results of the JV replications and could not be aggregated.

  In the UPV replication, the subjects take less time than in the JV replications. Initially, we were unable to find a cause for this discrepancy with respect to the JV replications. Again, we thought that it could be put down to training. Worse training would result in the subjects applying the functional technique poorly. This hypothesis would also confirm the initial assumption concerning the low effectiveness of the functional technique.

  The meeting helped to uncover the cause for the behaviour of the functional technique, which was not only related to training, but also to the low motivation of the subjects during the experiment. It was discovered that the subjects were not equally motivated during the execution of the experiment, because while the JV replications served to pass or fail the course, the experiment had hardly any effect on the final grade in the case of the UPV replication. This factor meant that performance was worse and, because of the low motivation, subjects did not make an effort to get good results.

- The **time it takes to apply the structural technique is as expected**. This result *is coherent* with results of the JV replications and can be combined.

  This meant that the structural technique's poor behaviour as regards effectiveness was even more dismaying.

  During the meeting, it was discovered that there was a scheduling error, and, as a result, the subjects did not have time to complete the application of the structural technique. This means that we have to add the fact that subjects were working under time pressure to the training and motivation effects.

During this stage of the replication, we faced no problems. Moreover, from the meeting between the groups of researchers, we were able to identify three causes to explain the discrepancies between the results of the JV and UPV replications: training, motivation and time. One of these (training) had already been identified as a possible cause for the discrepancies between the results of the JV and KL replications and was merely confirmed here. However, the two other causes of discrepancy (motivation and time) had been overlooked until now and were identified thanks to brainstorming among the groups of researchers familiar with the context, setting and operation of the replications and who are, therefore, more likely to identify discrepancies between the two contexts or the two operations jointly.

5.4 Evaluation of the Information Transmission Instrument

Comparing the results of Table 2 with the results in Table 1, we can see the benefits of using a *Family RP* and an *ICM with Agile Collaboration*. It is striking that the problems that emerged during the UPV replication were concentrated solely and exclusively in the operation phase (although they were discovered at the aggregation meeting).

Additionally, it should be noted that, while there were divergences between the results of the UPV replication and the JV replications, we managed to isolate the sources of variability in all cases thanks to the communication between researchers. As mentioned in the introduction, as opposed to other fields, it is very difficult in SE to get a context that is exactly the same as in the experiment that is to be replicated, mainly because of the shortage of resources. A greater control of the context is, therefore, essential. The communication among experimenters helps to contain the possible impact of the context variables on the experiment. The communication between researchers during the definition, planning, interpretation and aggregation stages has doubtless contributed to a satisfactory aggregation of results that can identify a cause behind the divergence for divergent results. However, it was not possible to completely eliminate divergences, precisely because the UPV did not manage to accurately reproduce the operation of the experiment as planned by the UPM.

Another point that is clear from Table 2, however, is that there is still room for improvement in the *Family RP* described earlier. Likewise, the *ICM with Agile Collaboration* can be fine tuned by including another meeting during the phases prior to experiment operation.

6. PROPOSED INSTRUMENT FOR TRANSMITTING INFORMATION FOR REPLICATION

Recapitulating, the ITI that we consider best suited for the replication and later aggregation of experimental results is composed of a *Family RP* and an *ICM with Agile Collaboration* that tailors the proposals of other authors on several points.

The *Family RP* proposed here is composed of three parts. The first is a general part that reflects how the item of knowledge examined by the experiment is matured and fashioned by the aggregation of the results of the replications. We term this part **Knowledge Part**. The second is a general part that describes the baseline experiment, that is, the experiment to be performed were the replicating researchers able to unconstrainedly tailor their context to the experiment setting. We term this part **Experiment Part**. And the third is a specific part that describes each replication of the experiment. We term this part **Replications Part**. The contents of each part are:

- **Knowledge Part** (one per experiment):
  - *Goals* of the experiment.
  - *List of associated replications.*
  - *History of the experiment*. Comparative schema of all the replications of which the experiment is composed (see [19], for example). This provides an overview of the replications, from which it will be clear what changes have been made in each replication.
  - *Aggregation* of the results of the different replications.
- **Experiment Part** (one per experiment):
  - *Definition and planning* of the replication.
  - *Experiment operation*. This includes: Detailed script containing what the experimenter should or should not do when running the experiment; Instructions for data collection; Description of the permitted atmosphere during the replication.
  - *Instructions for data filtering*, should any sort of pre-processing be necessary before analysis.
  - *Experiment material.* Description of the material required for experiment operation. This includes, for any type of experiment: Special-purpose material for use to train subjects; and General-purpose instructions sheets for subjects. For the experiment described in this paper: specifications and source code containing faults, and source code without faults for the training and experiment programs, and solution for training programs Data collection forms.
  - *Data analysis material*. For the experiment we looked at here, the description of the faults in the training and experiment programs, alongside the failures they caused when run.
  - *Electronic material* required for experiment operation, ready for replicating researchers to print out.
- **Replications Part** (one per replication):
  - All the items listed in the Experiment Part.
  - *Data collected* during the replication.
  - *Data analysis*.

One might think that the information about the material needed to run the experiment is redundant in the Replications Part. However, this material can change throughout the experiment's lifetime, as things may be improved upon or defects corrected. In this respect, the Experimental Part contains the baseline of the experiment. That is, the most up-to-date set of documentation needed for the operation of the experiment. If the experimental material is not included in each replication, there is a risk of the material used in each replication being lost if the material contained in the Experiment Part is modified. The aim here is to solve configuration management problems associated with the evolution of the experimental material due to improvement.

If any of the contents of the Replications Part are different from the Experiment Part, the changes made should be appropriately justified.

The RP proposed here is aimed at easing the aggregation of the results of different replications of one and the same experiment. It improves the RPs described in section 2 on several points.

On the one hand, it separates the evolution of the items of knowledge associated with the experiment from the information for running the experiment and from the particular information pertaining to each replication. This eases the job of the replicating researchers who can clearly identify (and record) their replication from the baseline experiment or from other replications.

On the other hand, the RP is self-contained insofar as it does not make use of external documents, like publications about the experiment, which in some cases may either not exist or perhaps impose constraints on the length or content of the document. We think it is vital for the RP to be self-contained because this makes it easier to locate all the replications of an experiment. Additionally, it is an aid for experimenters who do not have to search for and work from a lot of different documents.

Finally, the proposed structure for the RP solves the problems of configuration management in the RP, as it provides, on the one hand, the most up-to-date baseline of all the material available for the experiment operation and, on the other, makes a distinction between different versions of this material, tailored to the operation of a particular replication.

Apart from the structural question, this RP includes two completely new items: the schema for comparing replications and the data analysis for each replication. It also further specifies the item detailing the experiment operation, including a description of the atmosphere and the instructions on how subjects should fill in data collection forms.

Regarding the ICM proposed here, *Agile Collaboration* consists of the following procedure:

- **Kick-off Meeting**. This is held after the replicating researchers have examined the RP (at which point it does not yet contain a replication part for their replication). The goal of this meeting is to examine the context of the new replication, compare its setting and resources with what are required by the experiment and outline an adaptation that, while respecting the objectives of the experiment, matches the constraints of the new setting. At this meeting, any changes that need to be made to the experiment are decided upon jointly by the two groups of researchers with a view to designing the new replication.
- **Pre-Execution Meeting**. This is a second meeting in which the replicating researchers would bring up any doubts they had regarding the running of the experiment. It is held after appropriately updating the RP with the material related to the new replication
- **Visit** by one of the earlier experimenters during the running of the new replication to assure that it is performed as closely as possible to earlier replications. If this were not possible, a recording of the running of the experiment or a transcription of this recording could be sent to the replicating researchers, or one of the replicating researchers could visit the earlier experimenters during one of the experimental runs (in this order of preference).
- **Aggregation Meeting** to jointly interpret the results of the replications and look for the causes of convergences and divergences in results. In some replications described in the literature ([3]), replicating researchers take advantage of the chance of attending and observing while the earlier researchers run the experiment. Our proposal differs on this point, since our

experiences have led us to believe that the replicating researchers are obliged in this case to lend their attention to too many details, without knowing exactly which will and which will not be relevant for their replication.

The ICM proposed here is halfway between Occasional Collaboration [3], [17] and Heavy Collaboration [19]. It improves upon Occasional Collaboration in that it establishes permanent channels of communication between researchers. On the other hand, it is especially suited for environments in which Heavy Collaboration cannot be used. Researchers cannot always provide for such close collaboration (over several years), in which research groups have already worked together and new contexts are flexible enough so as not to require changes in the running of the experiment. Our proposal reduces the amount of communication between researchers to a minimum, while at the same time trying to assure the success of the replication. This results in a fewer meetings being held and collaboration being shorter term (six months in the case of the UPV replication).

7. CONCLUSIONS

The improvement in the ITI between researchers with the aim of enabling the accurate reproduction of experiments is a key issue in experimental SE. In this paper, we have presented our experiences in running several replications of an experiment with different ITIs. We have found that it is vital for the RP to be as comprehensive as possible to attain an accurate reproduction. Nevertheless, such a package is insufficient on its own. Inter-researchers Communication Mechanisms should be set up among experimenters to exchange information that is difficult to set out in document form, while at the same time not demanding too thorough a collaboration, which is often not feasible.

More specifically, the ITI we propose is composed of:

- A *Family RP*, which includes a knowledge part to reflect the evolution of the items of knowledge, an experiment part that contains the baseline to run the experiment and a replications part that is specific for each replication. The RP is self-contained, and there is no need to consult other documents to search for information about the experiment.
- An *ICM with Agile Collaboration* that includes two pre-experimental meetings, and one post-experimental meeting, in addition to visits during experiment operation.

Our proposal aims to improve the replication of experiments in SE. The proposed RP makes it easier for replicating researchers to run replications and aggregate their results with the knowledge gained from successive replications of the experiment. The proposed ICM establishes channels of communication among researchers to help them to perform an exact replication, without being overly demanding.

REFERENCES

[1] V.R. Basili, R.W. Selby. Comparing the Effectiveness of Software Testing Strategies. Department of Computer Science. University of Maryland. Technical Report TR-1501. College Park. May 1985.

[2] V.R. Basili, R.W. Selby. Comparing the Effectiveness of Software Testing Strategies. *IEEE Transactions on Software Engineering*. Pages 1278-1296. SE-13 (12), 1987.

[3] M. Ciolkowski, C. Differding, O. Laitenberger, J. Muench. Empirical Investigation of Perspective-Based Reading: A Replicated Experiment. ISERN Tech Report. ISERN-97-13. 1997.

[4] R. Conradi, V.R. Basili, J. Carver, F. Shull, G.H. Travassos. A Pragmatic Documents Standard for an Experience Library: Roles, Documents, Contents and Structure. University of Maryland Technical Report. CS-TR-4235. 2001.

[5] M. J¢rgensen. A Review of Studies on Expert Estimation of Software Development Effort. *Journal of Systems and Software.* 70 (1-2). Pp 37-60. 2004

[6] C.M. Judd, E.R. Smith, L.H. Kidder. *Research Methods in Social Relations.* Hartcourt Brace Jovanovich College Publishers. Orlando, Florida. 1991.

[7] N. Juristo, A.M. Moreno. *Basics of Software Engineering Experimentation*. Kluwer. 2001.

[8] N. Juristo, A.M. Moreno, S. Vegas. Reviewing 25 Years of Testing Technique Experiments. *Empirical Software Engineering*. Vol 9, N. 1, pages 7-44, 2004.

[9] N. Juristo, S. Vegas. Functional testing, structural testing and code reading: What fault type do they each detect? *Empirical Methods and Studies in Software Engineering- Experiences from ESERNET.* Springer-Verlag. Volume 2785. Chapter 12. Pages 235-261.2003.

[10] E. Kamsties, C. Lott. An empirical evaluation of three defect detection techniques. Technical Report ISERN 95-02, Dept. Computer Science, University of Kaiserslautern, May 1995.

[11] E. Kamsties, C.M. Lott. An Empirical Evaluation of Three Defect-Detection Techniques. *Proceedings of the Fifth European Software Engineering Conference*. Sitges, Spain. September 1995.

[12] O. Laitenberger, H.D. Rombach. (Quasi-)Experimental Studies in Industrial Settings. *Empirial Software Engineering*. Chapter 5. pp 167-227. World Scientific. 2003.

[13] J. Miller. Applying Meta-analytical Procedures to Software Engineering Experiments. *Journal of Systems and Software.* 54 (1). Pp 29-39. 2000.

[14] L.M. Pickard,, B.A. Kitchenham, P.W. Jones. Combining Empirical Results in Software Engineering. *Information and Software Technology.* 40 (14) pp 811-821. 1998.

[15] A. Porter, P. Johnson. Assessing software review meetings: results of a comparative analysis of two experimental studies. *IEEE Transactions on Software Engineering.* 23 (3). Pp 129-145. 1997.

[16] M. Roper, M. Wood, J. Miller. An empirical evaluation of defect detection techniques. *Information and Software Technology*. Vol. 39, pp 763-775. 1997.

[17] F. Shull, J. Carver, G.H. Travassos, J.C. Maldonado, R. Conradi, V.R. Basili. Replicated Studies: Building a Body of Knowledge about Software Reading Techniques. *Empirical Software Engineering*. Chapter 2. Pp 39-84. World Scientific. 2003.

[18] F. Shull, M. Mendoça, V. Basili, J. Carver, J.C. Maldonado, S. Fabbri, G.H. Travassos, M.C. Ferreira,. Knowledge-Sharing Issues in Experimental Software Engineering. *Empirical Software Engineering*. Vol 9 (1-2), pp 111-137. 2004.

[19] S. Vegas. Maduración de Conocimiento Mediante una Familia de Experimentos. *Jornadas Iberoamericanas en Ingeniería del Software Ingeniería del Conocimiento (JIISIC'04)*. Pp 487-500. An English version is available at http://is.ls.fi.upm.es/udis/miembros/sira/cv.html

[20] M. Wood, M. Roper, A. Brooks, J. Miller. Comparing and Combining Software Defect Detection Techniques: A Replicated Empirical Study. *Proceedings of the 6th European Software Engineering Conference*. Zurich, Switzerland. September 1997.