A Systematic Mapping Study on Testing Technique Experiments: Has the Situation Changed since 2000?

Jorge E. González[1], Natalia Juristo[2,3], Sira Vegas[2]
[1] Facultad de Ingeniería, Universidad Icesi, Colombia
[2] Facultad de Informática, Universidad Politécnica de Madrid, Madrid, Spain
[3] Department of Information Processing Science, University of Oulu, Oulu, Finland

ABSTRACT

Context: Juristo *et al.* [7] published a literature review about testing technique experiments. The goal was to provide a picture of which techniques and aspects of techniques had been studied experimentally, and try to compile a body of knowledge on testing techniques. *Goal:* In this paper, we extend Juristo *et al.*'s study to cover the years from 2000 (where it ended) until 2013. *Method:* We have performed a systematic mapping study. *Results:* The situation in testing experimentation has not changed since Juristo *et al.'s* study. *Conclusions:* The research field has the same shortcomings.

1. INTRODUCTION

In [7], Juristo *et al.* report a literature review on testing technique experiments. Its goals were to: (1) compile the body of knowledge on testing techniques and its maturity level; (2) provide a picture of which aspects of the techniques have been studied empirically, which need more testing, and which have not been considered. Their conclusions are: (1) many studies analyze results not using statistical techniques; (2) the main interest lies in the study of the techniques as separate from human influence: only a few studies use humans as experimental subjects in their experiments; (3) the interest focuses primarily on white box techniques; (4) there are few experiments using large and/or real programs; and (5) the techniques and the response variables examined vary enormously, and results cannot be aggregated.

The aim of this paper is to extend the review published by Juristo *et al.* in [7] to cover the period from 2000 to the present in order to check whether there has been any change in the trends regarding testing technique experiments. We have conducted a systematic mapping study in conformance with [8]. Mapping studies categorize primary studies to give an overview of the field of research under consideration.

The paper is organized as follows. Section 2 presents our goal and research questions. Section 3 details the search and selection process followed to identify the primary studies. Section 4 shows the data extracted from the relevant papers. Section 5 reports the results found. Finally, Section 6 presents our conclusions.

2. GOAL AND RESEARCH QUESTIONS

This research aims at summarizing the current state of the art in testing technique experiments with a view to a future compilation of the body of knowledge. We look for answers to the questions:

- RQ1. How many papers from 2000 to (May) 2013 report experiments on testing techniques?
- RQ2: What is the reporting coverage of the studies regarding the topics that should be covered according to published guidelines?
- RQ3: Do experiments use people as experimental subjects?
- RQ4. Which testing techniques and families of techniques have been studied experimentally?
- RQ5: Which aspects and response variables have the experiments examined?
- RQ6: What are the characteristics of the artifacts (programs) that have been used in the experiments?
- RQ7: Is there enough overlap among experiments so that techniques can be compared?

3. SEARCH AND SELECTION PROCESS

The **selected database** is SCOPUS™, as recommended in Dieste *et al.* [1]. The **scope** of our study is to continue Juristo *et al.'s* research [7], which studied the knowledge maturity of testing techniques experiments until year 2000. We plan to analyze experiments from 2000 to (May) 2013. However, some SLRs about regression testing techniques are identified [2], [3], [9], as well as a survey of mutation testing techniques [4]. We consider it unnecessary to include the regression testing and mutation testing technique families in our study, as they have already been covered by other reviews/surveys.

Two terms are chosen for the **search string definition**: TERM_A (*experiment*) and TERM _B (*testing technique*). We use the following inclusion criteria to further specify the search string: publication year greater than 1999; papers must belong to computer science and engineering areas; papers must be written in the English language.

Following the recommendations of Dieste *et al.* [1] we include in TERM_A: *experiment*, *experimental study*, *experimental comparison*, *experimental analysis*, *experimental evidence*, *experimental setting*, *experimentation*, *empirical study* and *empirical evaluation*. We also include: *experimental data*, *comparative*, *empirical research* and *empirical comparison*.

After reviewing [7] and some technical papers about testing techniques, we decide to include in TERM _B: *testing technique*, *test design method*, *test design technique*, *test case design technique*, *test case generation*, *test suite generation*, *random testing*, *functional testing*, *control flow testing*, *data flow testing*

and *improvement testing*.

Searching in titles, abstracts and keywords, we identify 937 relevant papers. Duplicates and irrelevant papers are excluded by analyzing titles and abstracts, remaining 91 papers. Later, the full document is analyzed, and 22 candidate papers are identified.

To study the threat to validity originated by the use of SCOPUS™ as only library, we used the three SLRs about regression testing techniques [2], [3], [9] as our gold standard. We compared the publications used in these studies with the list of relevant papers output by our search. Only 2 were not included in SCOPUS™ (6.5%). This suggests that our results can be trusted.

4. DATA EXTRACTION

We extract the following information from each paper[1]: publication type, topics reported, subjects used, testing techniques and family studied, response variables and aspects studied, and programs used.

The scope of our study does not cover case studies, surveys, experiments on mutation and regression testing techniques, and papers reporting the experiment as just part of the paper. We discard 8 out of the 22 relevant papers identified at the end of the search and selection process. [S3], [S22] and [S6] are excluded because they compare testing techniques with code review. [S19] and [S21] are excluded because they introduce novel testing technique approaches, and this research is confined to well-established techniques. [S1] is excluded because it presents theoretical results. Finally, [S13] and [S16] are omitted for RQ3-RQ7 because they present the same experiments that are described in [S14] and [S17] respectively. The final paper count is 16 for RQ1-RQ2 and 14 for RQ3-RQ7.

5. RESULTS

5.1 RQ1: Papers Reporting the Experiments

The left side of Figure 1 shows the number of papers that we have identified by type and year. There are 10 conference papers, 5 journal articles and 1 book chapter. The right side of Figure 1 shows the evolution of software testing technique experiments in terms of number of experiments performed per year. 62.5% (10/16) of the experiments were published from 2004 to 2008, although there is a gap in 2005 when no experiments were run. Only 3 experiments were run in the last five years (2009-2013). The number of experiments on testing techniques has declined over the last few years. This trend is opposite to the generally upward trend in the number of experiments run in SE. We have not been able to come up with an explanation for this.

5.2 RQ2: Reporting Coverage

We identify several *main* topics that an experiment should report, taking into consideration good reporting practices, along the lines presented in [5], [6], [10]. Figure 2 shows the number of primary studies that cover a given *main* topic. In general, the majority of the papers cover most of the *main* topics. However, there are 4 *main* topics (hypothesis, experimental subjects, origin of programs and statistical analysis) which are covered by only a few (from 2 to 5) studies. Statistical analysis has a direct bearing on the possibility of aggregating results. Use of experimental subjects, as well as the origin of the programs, has a direct bearing on the representativeness that results have of reality.

1 The list of revlevant papers can be found in http://www.grise.upm.es/sites/extras/9

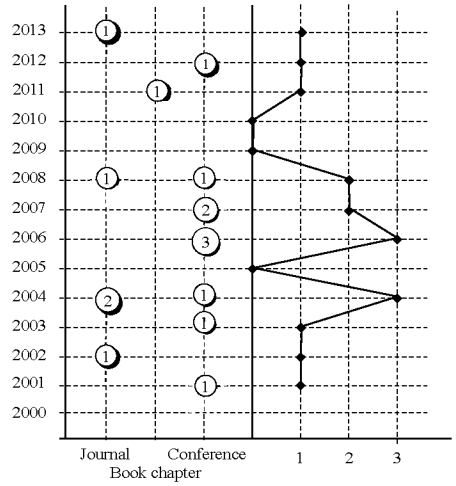

Figure 1. Evolution of relevant papers by type.

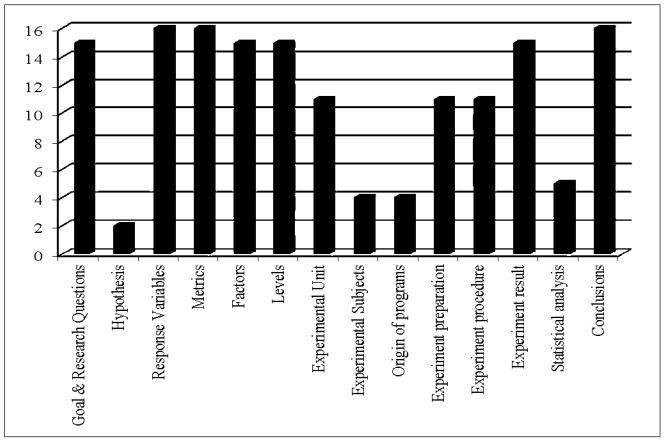

Figure 2. Coverage by *main* topics.

To evaluate the extent to which the papers cover the *main* topics that an experiment should report, we count how many *main* topics a given paper covers. Scores range from 6 to 13 (the perfect report covering all 13 main topics), and we are able to classify the papers according to three coverage levels: *low coverage papers* (scores 6-8), *medium coverage papers* (scores 9-11), and *high coverage papers* (scores 12-13). We identify 5 low-coverage reports, 9 medium-coverage reports and 2 high-coverage reports.

Figure 3 shows the evolution of experiments reporting coverage over time. We expect that experiment reporting increases over time, but we find that this is not the case. Moreover, there is only one study that achieves maximum quality coverage.

5.3 RQ3: Use of People

Only 14.3% (2/14) of the primary studies involve people. [S2] uses 300 experimental subjects, and [S10] uses 79 subjects. This means that almost all the identified experiments are interested in the pure application of techniques and do not take into account the influence of the human factor on technique behavior.

5.4 RQ4: Testing Techniques Studied

Columns 1-3 of Table 1 show the techniques examined in the experiments. Figure 4 shows the historical development chain, where the testing techniques and their family are related to each other through primary studies over time. The plotting of these data is modelled on the ideas of the schema used by Engstrom *et al.* [2]. This chart orders the primary studies (represented as

diamonds) identified in this research chronologically. If a study explores techniques belonging to one family, it will be positioned in the space reserved for that family. If it explores techniques that belong to two families, it will be positioned on the vertical dashed line between the two families. If the study cannot be positioned between two families ([S11]), the node representing the study in each of the respective family spaces is duplicated, and a dotted black line indicates the relationship.

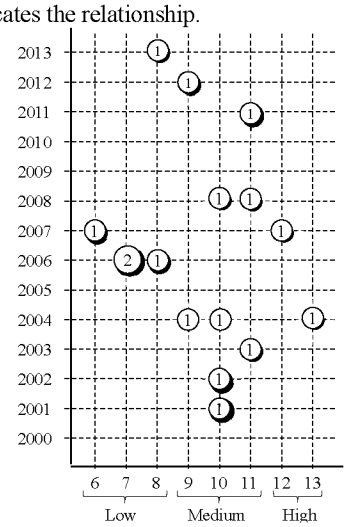

Figure 3. Evolution of experiments reporting coverage.

Table 1. Techniques examined in the primary studies.

Family	Id	Testing Technique	Studied
Functional Testing	T1	n-way (Combinatorial)	3
	T2	Equivalence Partitioning	1
	T3	Exploratory (ad hoc) Testing	1
	T4	Anti-Random Testing	1
	T5	Specification-Based Testing	1
Random Testing	T7	Pure Random	6
	T8	Adaptive Random Testing	3
	T9	RanGen	1
	T10	SymGen	1
Control Flow Testing	T11	Statement (Block) Coverage	1
	T12	Rule Coverage	1
	T13	Rule/Constraint Coverage	1
	T14	Decision (Branch) Coverage	2
	T15	Condition (Predicate) Coverage	3
	T16	Decision/Condition Coverage	2
	T17	Path Coverage	1
	T18	Loop Coverage	1
	T19	Boolean Operator Coverage	1
	T20	Relational Operator Coverage	1
	T21	Full Predicate Coverage	1
	T22	Modified Condition/Decision Coverage	1
	T23	Reinforced Condition/Decision Coverage	1
	T24	Round Trip Path Coverage	1
	T25	Disjunct Coverage	1
Data Flow Testing	T26	Automatic Test-Data Generation	1

Figure 4 also plots paths between the primary studies, where a path is a straight line linking studies that explore the same technique. Each path represents the studies that explore a given technique over time. The path length indicates how many studies examined a given technique. If two paths overlap partially or totally, they are distinguished by means of lines with different plots (solid or dashed). Paths are labelled with the name of the technique they represent, where the study representing the beginning of the path has been allocated. From Figure 4, we can conclude that:

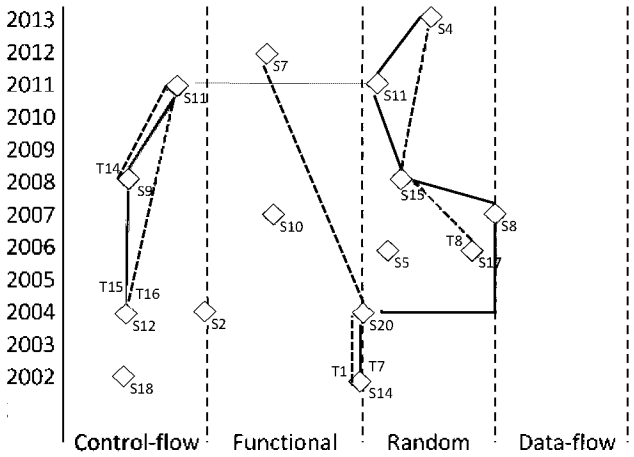

Figure 4. Testing technique families related to each other through primary studies.

- Most studies (9) compare techniques belonging to the same family: 2 functional, 4 random and 3 control flow.
- Only 5 studies compare techniques belonging to different families: 1 functional vs. control, 2 functional vs. random, 1 data flow vs. random, 1 control flow vs. random.
- More than 57.1% (8/14) of the studies focus on random testing techniques.
- None of the techniques explored by four studies are evaluated in any other study.
- There are 6 paths for the techniques: n-way (T1), pure random (T7), adaptive random (T8), decision coverage (T14), condition coverage (T15) and decision/condition coverage (T16).
- The most explored techniques are random: pure random (6 studies), and adaptive random testing (3 studies). These are followed by the n-way functional testing technique (3 studies). Finally, some control flow techniques are covered by more than one study: decision coverage (2 studies), condition coverage (3 studies), and decision/condition coverage (2 studies).

The last column of Table 1 shows how many primary studies explore a given technique. The overlap among studies is very low for all families. Globally, 77% (20/26) of the techniques are explored by only one primary study. Of the random techniques, 50% (2/4) are each explored by one study. The only data flow technique is examined by just one study. 80% (12/15) of the control flow techniques are investigated by only one study. Finally, 80% (4/5) of the functional testing techniques are investigated by only one study.

5.5 RQ5: Response Variables Used

Table 2 shows the aspects and response variables examined by each study. We can see that there is hardly any overlap among response variables. 81% (17/21) of the response variables are used by one study. The only exception is the *number of test cases*, which is used by 23.8% (5/ 21) studies. This will make it difficult to compare the results of the different studies.

5.6 RQ7: Characteristics of the Artifacts Used

Table 3 shows the characteristics of the programs used. Empty cells mean that the study does not contain the information. 93% (13/14) of the primary studies detail the number of programs used, and 57% (8/14) specify their size. 77% (10/13) of the primary studies use just few programs (up to 5); 23% (3/13) use 9 or more

programs. Most of the studies (62%) use medium-sized programs (5/8); 12.5% (1/8) tiny programs, and 25% (2/8) large programs.

Table 2. Response variables used by each study.

Aspect	Id	Response variable	Primary study
Effectiveness	RV1	Mutation score	[S4] [S11] [S14]
	RV2	Block coverage	[S4]
	RV3	Decision coverage	[S4]
	RV4	C-uses coverage	[S4]
	RV5	Detected faults	[S7] [S20] [S12]
	RV6	F-measure	[S17] [S15] [S18]
	RV7	E-measure	[S15]
	RV8	P-measure	[S17]
	RV9	Fraction of faults revealed	[S9]
	RV10	Coverage	[S8]
	RV11	Number of non-arithmetic-overflow test generated (TG)	[S5]
	RV12	Number of failing tests (FT)	[S5]
Efficiency	RV13	Number of test cases	[S7] [S11] [S9] [S8] [S18]
	RV14	Search time to generate test suite	[S8]
	RV15	Number of iterations for generating test suite	[S8]
	RV16	Number of defects found during time	[S10]
Precision	RV17	Ratio of fault-revealing tests (FT/TG)	[S5]
Reliability	RV18	S-measure	[S15]
Cost-effectiveness	RV19	Number mutants killed/ number of test cases	[S2]
	RV20	Number mutants killed/ amount of CPU consumed	[S2]
	RV21	Number of mutants killed/ number LOC test drivers	[S2]

Table 3. Characteristics of artifacts used.

Study	Number of programs	Size (LOC)
[S2]	4	145-495
[S4]	10	90-842
[S5]	16	21-1832
[S7]	2	5,381-11803
[S8]	9	20-61
[S9]	5	310-1490
[S10]	1	-
[S11]	4	-
[S12]	-	-
[S14]	1	-
[S15]	2	-
[S17]	2	-
[S18]	1	653
[S20]	2	6,200-8,700

5.7 RQ8: Can Techniques be Compared?

Figure 5 (as in [2]) gives an overview of the techniques that have been studied. Circles represent techniques and lines connecting circles represent the number of studies in which they have been compared. Numbers on the lines represent the number of studies in which techniques are compared. To simplify the diagram, lines without a number represent techniques examined in one study.

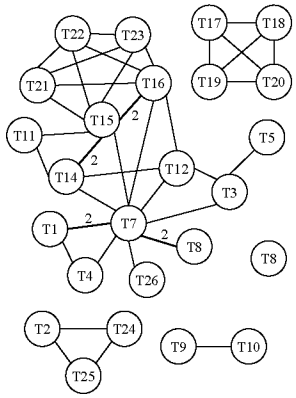

Figure 5. Techniques compared to each other by experiments.

Only four pairs of techniques are candidates to be compared: T1-T7, T7-T8, T14-T15, and T15-T16, as each pair of techniques has been evaluated by at least 2 primary studies. However, comparison can only be done if the studies use the same response variable. Table 4 shows the response variables measured in the studies comparing the four pairs of techniques. We find that only T14 and T15 can be compared, as [S9] and [S11] share RV13.

Table 4. Response variables, studies and techniques.

Techniques	Study	Response variable
T1-17	[S14]	RV1
	[S20]	RV5
T7-T8	[S4]	RV1, RV2, RV3, RV4
	[S15]	RV6, RV7, RV18
T14-T15	[S9]	RV9, RV13
	[S11]	RV1, RV13
T15-T16	[S11]	RV1, RV13
	[S12]	RV5

6. CONCLUSIONS

We have performed a systematic mapping study on testing technique experiments, extending Juristo *et al.*'s study [7] to cover the past 13 years. The goal of the research was to find out if it would be possible to combine the results of the experiments to discover testing technique knowledge. At the same time, we wanted to know whether the testing technique experimentation picture had changed since Juristo *et al.*'s study.

The results suggest that the testing technique experimentation panorama has not changed much over since Juristo *et al.'s* study [7]. Experiments are few and far between. They are not realistic as they do not consider the application of the techniques by people. They do not usually report statistical analysis (which places a constraint on any combination of results). There is still a shortage of replications for combining results, and the response variables used differ, which is an obstacle to the combination of results.

7. ACKNOWLEDGMENTS

This work was supported by Spanish Ministry of Economy and Competitiveness research grant TIN2011-23216.

8. REFERENCES

[1] O. Dieste, A. Grimán, N. Juristo. Developing search strategies for detecting relevant experiments. *Empirical Software Engineering*. vol 14, no. 5, p. 513–539, 2009.

[2] E. Engstrom, P. Runeson, M. Skoglund, A systematic review on regression test selection techniques, *Information and Software Technology*, vol. 52, no. 1, p. 14 – 30, 2010.

[3] E. Engstrom, M. Skoglund, P. Runeson, Empirical evaluations of regression test selection techniques: A systematic review, In Proceedings of the 2nd *ESEM*, pp. 22-31, 2008.

[4] Y. Jia, M. Harman, An analysis and survey of the development of mutation testing, *IEEE Transactions on Software Engineering*, vol. 37, no. 5, p. 649 – 678, 2011.

[5] A. Jedlitschka, M. Ciolkowski, D. Pfahl. Reporting Experiments in Software Engineering. *Guide to Advanced Empirical Software Engineering*. Springer, 2008.

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

[7] N. Juristo, A.M. Moreno, S. Vegas, Reviewing 25 Years of Testing Technique Experiments. *Empirical Software Engineering*, vol. 9, no. 1-2, pp. 7-44, 2004.

[8] K. Petersen, R. Feldt, S. Mujtaba, M. Mattsson. Systematic mapping studies in software engineering. In Proceedings of the 12th *EASE*. pp. 71–80, 2008,

[9] Y. Singh, A. Kaur, B. Suri, S. Singhal, Systematic literature review on regression test prioritization techniques, *Informatica*, vol. 36, no. 4, p. 379 – 408, 2012.

[10] C. Wohlin, P. Runeson, M. Höst. *Experimentation in Software Engineering*. Springer, 2012.