# Analysis of Inspection Technique Performance

O. Dieste, E. Fernández, P. Pesado, R. García-Martínez

Grupo de Ingeniería de Software Experimental. Facultad de Informática. UPM
Programa de Doctorado en Ciencias Informáticas. Facultad de Informática. UNLP
Instituto de Investigaciones en Informática LIDI. Facultad de Informática. UNLP - CIC
Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. UBA
Área Ingeniería del Software. Licenciatura en Sistemas. Departamento Desarrollo Productivo y Tecnológico. UNLa.
odieste@fi.upm.es, qiqe2000@hotmail.com, ppesado@lidi.info.unlp.edu.ar, rgarciamar@fi.uba.ar

**Abstract.** Inspection techniques are strategies for analysing software artefacts. These techniques provide guidelines for examining the software documentation and identifying defects. These guidelines consist of a series of heuristics to help reviewers to read and understand the artefact that they are analysing. A number of researchers have now developed experimental studies to compare the performance of the different techniques in an attempt to find out what is the best strategy to adopt in which cases. In this work, we conduct a systematic review of the effectiveness and efficiency of inspection techniques following Kitchenham's recommendations.

**Keywords.** Inspection technique. Systematic review. Techniques Performance. Meta-analysis. Experimental Software Engineering.

## 1. Introduction

Inspection techniques [1] are strategies for analysing software artefacts (requirements specification, design specifications, source code, etc.) Fagan [2] originally proposed this type of technique in 1976 as an early error detection method for information systems. Since then substantial effort has been put into optimizing the inspection process [3]. This optimization focused on several aspects of the process, such as the number of inspectors, the optimum reading rate and the optimum size of the artefact for inspection, etc. According to Denger [1], the most important step in each inspection is the defect detection phase, where inspectors individually try to identify as many defects as possible in the artefact under inspection. Reading techniques were defined to optimize defect detection. Reading techniques help inspectors to identify more defects with less effort. This makes the inspection process more efficient.
Following on from this, a number of inspection techniques were invented. These techniques provide guidelines for examining the software documentation and identifying defects. These guidelines consist of a series of heuristics to help reviewers to read and understand the artefact that they are analysing. They were developed to improve the performance of what are known as ad hoc techniques. The best known

reading techniques are [1, 4]: checklist-based and scenario-based techniques. The scenario-based techniques are further divided into perspective-based, use-based and defect-based reading techniques. A number of researchers have now developed experimental studies to compare the performance of the different testing techniques in an attempt to find out what is the best strategy to adopt in which cases. Also they have developed surveys comparing testing techniques with inspection techniques [5]. Yet nobody has so far conducted a study to compare the performance of different reading techniques with each other. Additionally, the above survey papers are qualitative (closely connected to the researcher's opinion), even though there are enough existing studies to conduct a quantitative survey based on meta-analysis. This would make the evidence gathered from the aggregation process more reliable.

In this paper, then, we conduct a systematic review of the effectiveness and efficiency of inspection techniques following Kitchenham's recommendations [6]. To do this, the paper is structured as follows. Section 2 describes how the different inspection techniques work. Section 3 describes the systematic review that we conducted. Section 4 discusses the results, and Section 5 describes the conclusions.

## 2. State of the Art

The early inspection techniques did not set out any strategy or special guidance for going about the task of inspecting an artefact, and an inspector's skill and experience had a considerable impact on the number of detected defects [7]. These techniques are known as an ad hoc techniques, and they were very often opposed on the grounds of low dependability. To improve this aspect, a set of strategies were developed (based on guidelines indicating how to do the reading). These strategies were intended to the make the reviewers' work more dependable and less reliant on their experience. The most used strategies are [1, 4]: checklist-based and scenario-based inspection techniques. These techniques are described below:

- The checklist-based reading (CBR) techniques [1, 8] provide inspectors with a list of questions about potential defects in the inspected artefact. This way, CBR provides guidelines about what to look for during defect detection. However, there are no guidelines about how to look for defects, that is, there are no heuristics instructing inspectors how to find out whether they should answer a checklist question yes or no. Also, each inspector has to verify the whole artefact against all the checklist questions. Bearing in mind the many questions that have to be addressed, this can be an extremely onerous undertaking in complex artefacts.
- Scenario-based reading (SBR) techniques[1] were designed with the aim of overcoming the weaknesses of checklist-based reading: 1) provide guidance about how to actively develop the inspection on the artefact, which is called *active guidance*; and 2) confine the reviewer's focus to one specific aspect of interest, that is, indicate what to inspect, which is called *separation of concerns.*

  These concepts of the scenario-based approach are implemented as "reading scenarios". A reading scenario is basically composed of three parts: 1) an introduction that defines the focus of a scenario; 2) a series of instructions providing a step-by-step description of the activities that inspectors should carry out to detect defects; and 3) questions that focus inspectors on what quality aspects to look for as they carry out the instructions.

SBR techniques are further divided into defect-based techniques, perspective-based techniques and usage-based techniques. They are described below:

- Perspective-based reading (PBR) techniques [9] provide guidance to help inspectors to adopt the standpoints of the key stakeholders of the item under inspection. The key perspectives are defined as: designer (D), tester (T) and end user (U). Inspectors create abstractions that are relevant for the perspective they are dealing with. For example, a designer creates preliminary high-level diagrams, a tester creates a series of test cases and a user creates a series of use cases. As they create the abstractions, inspectors use a series of questions to help to detect defects. The questions are generally based on a common defect classification. This is not a static set of defect types and can be adapted as necessary to each setting.
- Usage-based reading (UBR) techniques [10] focus on the quality of the product from the user's viewpoint. UBR is underpinned by use cases and set use case priorities. Use cases are used to guide reviewers through the software item during inspection. Use cases are prioritized by order of importance depending on the developed system's user requirements. This way, the reviewers that use UBR focus first on the important parts, finding as a result the fault that match the defects that users consider to be most significant.
- Defect-based reading (DBR) techniques [4] are mainly applied to requirements documents. The main idea behind this technique is that different reviewers should focus on different defect classes as they inspect an item. To do this, they are applied by work groups. Although this helps to identify more defects, it also increases application costs because more human resources are required.

## 3. Systematic Review

A systematic review (SR) [6] is a method for identifying, evaluating and interpreting all research pertaining to a particular research question, subject area or phenomenon of interest. Although there are many grounds for developing a SR, the reasons in this case were two. First, although experiments comparing the performance of different inspection techniques are now being run, they are generally small (e.g. the number of subjects in [1, 11, 12, 13] is no more than 20). Therefore they do not provide evidence enough to be able to positively state whether one inspection technique is better (more effective and/or more efficient) than another. Second, the findings of the experiments run to date are not consistent (e.g. some experiments find that scenario-based techniques are better than checklist-based techniques, whereas others fail to find any significant difference between the two techniques, as mentioned in [14]). This places a serious constraint on the use of empirical knowledge in both academia and industry.

The SR process has been developed by dividing the work into two key stages: 1) definition of the research question, and 2) experiment search and selection and data extraction.

### 3.1. Definition of Research Questions

The goal of this paper is to determine which inspection strategy is best. In our view, the response variables that we should analyse to do this are effectiveness (linked to the number of errors each technique manages to identify) and efficiency (linked to the number of errors detected by time unit). Therefore, the research questions are:

1. Which types of inspection techniques (ad hoc, CBR or SBR) are most effective?
2. Which types of inspection techniques (ad hoc, CBR or SBR) are most efficient?

### 3.2. Experiment Search and Selection

To be able to guide the search process we set the following inclusion/exclusion criteria:

- The paper must compare the performance of two or more inspection techniques.
- The aspects for comparison are effectiveness and/or efficiency.
- The paper describes the setting in which the experiment is run: professional or academic environments.
- The paper was published in congress proceedings or a refereed journal.
- The paper uses experimental subject randomization.

The search process was divided into three stages. The first, preliminary, stage was developed on the Google search engine [15]. The second, more formal and structured search, was developed on the Scopus research literature database [16]. The third search was run on the bibliographic references in the selected studies. The inclusion/exclusion criteria were refined as the search progressed.

The search words, described in Table 1, were built on the basis of the research questions, the inclusion/exclusion criteria, the knowledge gathered in the preliminary Google search (where the terms "perspective-based reading", "checklist based reading", "scenario-based reading", "usage-based reading", "defect-based reading" were identified) and the suggestions in [17] (which led to the addition of the terms “experiment” “empirical study” “empirical study” to the search string):

**Table 1.** Search strategies.

| Source | Search fields | String used |
|---|---|---|
| Google | | “code inspection” “experiment” “empirical study” “empirical study” |
| SCOPUS™ | Title, abstract and keywords | TITLE-ABS-KEY(“perspective based reading” OR “checklist based reading” OR “scenario based reading” OR “usage based reading” OR “defect based reading”) AND (“experiment” OR “empirical study” OR “empirical study”) |

As a result of the search process we managed to identify 44 studies, of which 20 were selected. To do this, each study was analysed by two reviewers. These reviewers then reached agreement on whether or not the study should be included. Note that these studies often cover more than one experiment. We managed to identify a total of 48 experiments, of which 45 were run on senior students taking computing-related degrees and only three were run on professionals from the field. As regards the statistical parameters, eight experiments do not detail either the means or the variances and only specify whether one treatment was better than another, 28 experiments do not publish the variances (indicating just the number of subjects and means) and only 12 experiments publish all the statistical parameters (for further details about identified studies see [20]).

## 4. Results Synthesis

The quantitative synthesis of results involves combining the results of a set of previously selected experimental studies to get a single final result. The most popular method of synthesis (or meta-analysis as it is commonly referred to in the literature) is the weighted mean difference (WMD) [21]. To use this method, the experimental studies must publish the following statistical parameters: sample size, averages and variances or standard deviation. Unfortunately, many of the selected experimental studies do not publish all the statistical parameters. This prevents WMD from being used for aggregation.

For this reason, the results of the selected experiments will be combined using the response ratio (RR) method suggested by Lajeunesse and Forbes [22] and Friedrich and colleagues [23]. The RR involves estimating an effect index or ratio between an experimental and a control treatment using the quotient of both means (RR = $Y^E$ / $Y^C$). This quotient estimates how much better one treatment is than the other. So, for example, a ratio of 1.3 indicates that the experimental treatment is 30% better than the control treatment, whereas a ratio of 1 indicates that there is no difference between the performance of the two treatments.

To assure that the combination of a set of studies is more precise, the natural logarithm was added to the method. The natural logarithm can linearize the results and normalize their distribution. This makes it a good method for use to estimate effects when the set of experiments is small [13]. Note that after estimating the index, the anti-logarithm must be applied to the result to get the final effect index.

The method is applied in a two-step process. First we have to estimate the ratio of each experiment. Then, based on the ratios of all the experiments, we estimate the global ratio or effect by calculating a weighted average of the individual ratios, as shown in Eq. 1:

$$L^* = \frac{\sum_{i=1}^{k} W_i^* L_i}{\sum_{i=1}^{k} W_i^*} \tag{1}$$

$L^*$ is the global effect
$L_i$ is the effect of each study
$W_i$ is the weight factor = $1/v$

Note that the weight factor of the studies can be estimated by applying the *parametric* RR method (PRR), where the studies are weighted by the inverse variance (as indicated in Eq. 2) or the *non-parametric* RR method (NPRR), where the studies are weighted by the number of experimental subjects (as shown in Eq. 3). The main advantage of the non-parametric method is that the variances of the studies do not have to be known, whereas the key benefit of the parametric method is that it is extremely precise even if the studies include few subjects.

$$v = \frac{S^{2E}}{n^{E}Y^{E}} + \frac{S^{2C}}{n^{C}Y^{C}} \tag{2}$$

$v$ is the standard error
$S^2$ is the study variance
Y is the study mean
n is the number of subjects

$$v = \frac{n_C + n_E}{n_E n_C} + \frac{Ln(RR^2)}{2(n_C + n_E)} \tag{3}$$

$v$ is the standard error
n is the number of subjects
RR is the ratio

Having estimated the effect size, we can estimate its confidence interval as indicated in Eq. 4.

$$L^{*} - Z_{\alpha/2}\sqrt{v} \le \lambda \le L^{*} + Z_{\alpha/2}\sqrt{v} \tag{4}$$

Z is the number of standard deviations separating the mean from the endpoint at the specified significance level.

For more details about how to apply the above formulae, see [24].

Note that, as the studies that can be aggregated using PRR, can also be aggregated by NPRR, the studies that publish all the statistical parameters will be first combined using PRR and then lumped together with the studies that do not publish all the statistical parameters for combination using NPRR. This way, we expect to find out whether there are changes of trend in the results when incompletely reported studies are added. In the following, we summarize the results. The results were estimated at a confidence interval of 95%.

The Analysis of  Use was done by grouping the experiments by the analysed item (requirements or design reports), uses of techniques (individual or group) and the evaluated response variable. Table 2 shows, the estimated effect size (RR), plus its respective confidence interval (L_endp and U_endp) and the number of experiments aggregateds (K).

## 5. Discussion

The NPRR method proved to be a very useful method for applying meta-analysis in the field of inspection. We were able to combine 35 of the 48 identified experiments. If we were to have applied the weighted mean difference (WMD) method in groups of experiments greater than two, this figure would have been reduced to just three. This is because, to be able to estimate the effect size using NPRR, the experiments do not have to report variance as they do with the WMD and RR methods.

Even though a sizeable number of experiments are available, we gathered relatively little evidence, and what evidence we did get is not overly reliable. The main reason is the wide range of application contexts and factors covered by the primary studies. In actual fact, we have experiments that study the response variables (effectiveness and efficiency) at the individual or group level, using requirements or design documents and applying alternative SBR techniques (PBR, UBR and DBR). In view of this diversity, we were obliged to divide the original set of 35 experiments into more uniform subsets (to reduce the threat of heterogeneity), but, in return, these subsets contain fewer experiments, making the result of the meta-analysis less conclusive.

**Table 2.** Analysis of techniques applied individually.

| Group | Result | Observations |
|---|---|---|
| CBR vs PBR / Design / Individual / Effectiveness | RR = 1.047<br>L_endp = 0.991<br>U_endp = 1.106<br>K = 3 | This is the only case where paper were able to be aggregated by RRP.<br>CBR has a slight advantage over PBR. This would appear to be significant, as the confidence interval contains the value 1 by only one hundredth. |
| | RR = 0.977<br>L_endp = 0.991<br>U_endp = 1.106<br>K = 4 | The NPRR method had to be used to add the fourth experiment, as this experiment does not publish variances. With this addition, the apparent advantage of the CBR method dissolves. |
| CBR vs PBR / Requirements / Individual / Effectiveness | RR = 1.156<br>L_endp = 0.923<br>U_endp = 1.449<br>K = 5 | The CBR technique has an advantage over PBR. But the CBR technique cannot really be said to be better than PBR, because the value 1 is well inside the confidence interval. |
| CBR vs PBR / Requirements / Individual / Efficiency | RR = 0.847<br>L_endp = 0.550<br>U_endp = 1.304<br>K = 3 | The PBR technique has an advantage. Again, the PBR technique cannot really be said to be better than CBR because the value 1 is well inside the confidence interval. |
| CBR vs UBR / Design / Individual / Effectiveness | RR = 0.663<br>L_endp = 0.515<br>U_endp = 0.853<br>K = 3 | The UBR technique is 35% better than the CBR technique. This is backed by the fact that the confidence interval does not contain the value 1. Therefore, this difference can be said to be statistically significant. |
| CBR vs UBR / Design / Individual / Efficiency | RR = 0.753<br>L_endp = 0.548<br>U_endp = 1.036<br>K = 3 | The UBR technique is 25% better than the CBR technique: But, contrary to what applied to effectiveness, the confidence interval contains the value 1. Even so, it would not be overly venturesome to claim that the UBR technique is more efficient than the CBR technique, as the non-parametric method tends to return greater confidence intervals than normal if variances are not very high. |
| PBR vs Ad Hoc / Design / Individual / Effectiveness | RR = 1.141<br>L_endp = 0.657<br>U_endp = 1.983<br>K = 4 | The PBR technique has an advantage, but it cannot really be said to be better because the value 1 is well inside the confidence interval. |
| CBR vs PBR / Requirements / group / Effectiveness | RR = 0.977<br>L_endp = 0.627<br>U_endp = 1.523<br>K = 5 | Neither of the techniques can be said to be better than the other. The effect is more or less 1. |
| CBR vs PBR / Requirements / group / Efficiency | RR = 1.094<br>L_endp = 0.584<br>U_endp = 2.052<br>K = 3 | The CBR technique is 10% better than the PBR technique. But, as the confidence interval is very wide (with only four experimental studies, the three experiments are very small) and the value 1 is well inside, neither of the techniques can be said to be better than the other. |
| CBR vs Ad-Hoc / Requirements / group / Effectiveness | RR = 0.942<br>L_endp = 0.581<br>U_endp = 1.526<br>K = 6 | The ad hoc technique would appear to be 5% better than the CBR technique, but, as the value 1 is well inside the confidence interval, neither of the techniques can be said to be better than the other. |
| PBR vs Ad-Hoc / Design / group / Effectiveness | RR = 1.015<br>L_endp = 0.638<br>U_endp = 1.613<br>K = 6 | Neither of the techniques can be said to be better than the other. The effect is more or less 1. |

The synthesis has turned up several interesting results:

- The first is related to the CBR and PBR techniques. Used to inspect requirements documents, both the individual and group measurements showed the CBR technique to be equivalent to the PBR technique in terms of both effectiveness and efficiency. If used to inspect design documents, the situation is nowhere near as clear cut because there is very little evidence in this respect. Even so, what little evidence we have indicates that: (1) CBR is equivalent to PBR in terms of effectiveness measured individually; (2) the experiments analysing individual efficiency show the techniques to be clearly on a par; and (3) there are no data at all at group level. It would then be chancy to make any claim about the performance of these techniques as regards the analysis of design documents.

- The second interesting result refers to the relationship between the CBR and UBR techniques. In this case, we only have experiments run on design documents. However, UBR proved to perform better than the CBR technique on the two measured aspects: effectiveness at the individual level and efficiency at the individual level. This improvement is as much as 35% in the case of effectiveness and 25% in the case of efficiency (although it was not, in this case, statistically significant at $\alpha = 0.05$). We were unable to analyse the CBR and UBR techniques at group level because the two experiments that analyse their effectiveness do not publish the necessary information to be able to combine their results.
- The third noteworthy result is that we were unable to aggregate the experiments that we had identified comparing the CBR and DBR techniques using RR because their reports did not publish the treatment means.
- Finally, contrary to expectations, the ad hoc techniques tend to be more effective than the CBR technique and less effective than the PBR technique, but we were unable to find significant differences in any of the analysed cases. Nor do we know whether the reviewers that applied the ad hoc techniques always applied the same strategy or were equally experienced, these being aspects that are likely to influence the effectiveness of the ad hoc technique.

## 6. Validity Threats

Although we have arrived at some interesting findings, there are a number of concerns that could pose a threat to their validity:

- Application of the knowledge in industry: As only three of the 48 selected experiments were developed in industry (the others were run in academic settings), there is no guarantee that these findings are directly applicable to industry.
- Study diversity: Many of the identified experiments are replications of other earlier experiments. Whereas this helps to assure that the studies are homogeneous, it does restrict the field of application of the techniques. Consequently, it acts as a constraint on the generalization of the findings.
- Experiment size: We have identified a sizeable number of experiments, and they have been aggregated using meta-analysis, which helps to improve results reliability. On the other hand, though, most of the experiments are small, which accentuates the risk of experimental error.
- Power of the NPRR method: Even though the NPRR method is highly reliable for a non-parametric technique, it is a method that often tends to output false negatives when the experiment variance is medium or low. This can influence the findings somewhat, where differences in studies indicating that one technique is better than another tend to be classed as insignificant.

## 7. Conclusion

RR is a very versatile method for combining experiments, as suggested in research by Lajeunesse and Forbes [23] and Friedrich and colleagues [24]. They show that the PRR method is just as reliable and statistically powerful as WMD, especially when there are few experiments. NPRR, on the other hand, tends to be less statistically powerful. Even so, it behaved very well in this research, as it managed to turn up significant results on several occasions.

The meta-analysis developed showed that the performance of the CBR and PBR techniques is quite similar (in terms of both effectiveness and efficiency, used on requirements or design documents, and applied by individuals or a group). However, the addition of more experiments to the aggregation process is an absolute must if we are to come up with more evidence to justify the results.

On the other hand, we can say that the UBR technique is to be preferred to CBR. This has only been proved with respect to the effectiveness of the techniques applied to design documents reviewed individually. Outside this context, there is no statistical foundation to claim that UBR is better than CBR.

Finally, there is no significant evidence to claim that reviewers using heuristics-based techniques (like CBR and SBR) perform better than when they apply their own judgement.

## 8. References

1. Denger, C., Ciolkowski, M., Lanubile, F.; Does Active Guidance Improve Software Inspections? A Preliminary Empirical Study; 2004; Proceedings of the IASTED International Conference SOFTWARE ENGINEERING February 17-19, 2004, Innsbruck, Austria; 408-413
2. Fagan, M.E., Design and Code inspections to reduce errors in program development, 1976, IBM Systems Journal, Vol. 15, No 3, Page 182-211
3. Sabaliauskaite, G., Matsukawa, F., Kusumoto, S., Inoue, K.; 2002; An Experimental Comparison of Checklist-Based Reading and Perspective-Based Reading for UML Design Document Inspection; Proceedings of the 2002 International Symposium on Empirical Software Engineering (ISESE'02)
4. Berling, T., Thelin, T.; 2004; *A Case Study of Reading Techniques in a Software Company*; Proceedings of the 2004 International Symposium on Empirical Software Engineering (ISESE'04)
5. Runeson, P., Andersson , C., Thelin, T., Andrews, A., Berling, T.; 2006; What Do We Know about Defect Detection Methods?, IEEE Software, v.23 n.3, p.82-90, May 2006
6. Kitchenham, B. A.; 2004; *Procedures for performing systematic reviews*. Keele University; TR/SE-0401. Keele University Technical Report.
7. Sandahl, K., Blomkvist, O., Karlsson, J., Krysander, C., Lindvall, M., Ohlsson, N.; 1998; An Extended Replication of an Experiment for Assessing Methods for Software Requirements Inspections; 1998; Empirical Software Engineering, 3, 327–354
8. Basili, V. R., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sörumgård, S., Zelkowitz, M.; 1996; *The empirical investigation of perspective-based reading*, International Journal on Empirical Software Engineering, Vol. 1, No. 2; pp. 133–164.

9. Maldonado, J., Carver, J., Shull, F., Fabbri, S., Dória, E., Martimiano, L., Mendonça, M., Basili, V.; 2006; Perspective-Based Reading: A Replicated Experiment Focused on Individual Reviewer Effectiveness; Empir Software Eng (2006) 11: 119–142
10. Thelin, T., Andersson, C., Runeson, P., Dzamashvili-Fogelström, N.; 2004; A Replicated Experiment of Usage-Based and Checklist-Based Reading; Proceedings of the 10th International Symposium on Software Metrics (METRICS'04)
11. Thelin, T., Runeson, P., Wohlin, C.; 2003; An Experimental Comparison of Usage-Based and Checklist-Based Reading; IEEE Transactions on Software Engineering, Vol. 29, No. 8, August 2003; pp. 687-704
12. Porter, A., Votta, L.; 1998; Comparing Detection Methods For Software Requirements Inspections: A Replication Using Professional Subjects; 1998; Empirical Software Engineering, 3, 355–379 (1998)
13. Laitenberger, O., El Emam, K., Harbich, T.; 2001; *An Internally Replicated Quasi – Experimental Comparison of Checklist and Perspective – Based Reading of Code Documents*; IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 27, NO. 5, MAY 2001; 387-421
14. Lanubile, F., Mallardo, T., Calefato, F., Denger, C., Ciolkowski, M.; 2004; Assessing the Impact of Active Guidance for Defect Detection: A Replicated Experiment; Proceedings of the 10th International Symposium on Software Metrics (METRICS'04)
15. www.google.com
16. www.scopus.com
17. Dieste, O., Grimán Padua, A.; 2007, Developing Search Strategies for Detecting Relevant Experiments for Systematic Reviews, First International Symposium on Empirical Software Engineering and Measurement, 215-224
18. Biffl, S., Halling, M.; 2003; Investigating the Defect Detection Effectiveness and Cost Benefit of Nominal Inspection Teams; IEEE Transactions on Software Engineering, Vol. 29, No. 5, May 2003; pp. 385-397
19. Winkler, D., Halling, M., Biffl, S.; 2004; Investigating the Effect of Expert Ranking of Use Cases for Design Inspection; Proceedings of the 30th EUROMICRO Conference (EUROMICRO'04)
20. Malacrida, J.; 2009; Revisión y agregación de estudios experimentale vinculados a técnicas de inspección ; Tesis de Maestría en Ingeniería del Software; Laboratorio de Sistemas Inteligantes; Facultad de Ingeniría; Universidad de Buenos Aires
21. Hedges, L.; Olkin, I.; 1985; Statistical methods for meta-analysis. Academic Press
22. Lajeunesse, M & Forbes, M.; 2003; *Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques*. Ecology Letters, 6: 448-454.
23. Friedrich, J, Adhikari, N; Beyene, J; 2008; The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: A simulation study; BMC Medical Research Methodology
24. Hedges; L.; Gurevitch. J.; Curtis, P.; 1999; The Meta-Analysis of Response Ratios in Experimental Ecology. Ecology. 80(4), pp. 1150-1156