Report on The First International Workshop on Comparative Evaluation in Requirements Engineering

Vincenzo Gervasi
University of Pisa
gervasi@di.unipi.it

Didar Zowghi
University of Technology, Sydney
didar@it.uts.edu.au

Steve Easterbrook
University of Toronto
sme@cs.toronto.edu

Susan Elliott Sim
University of California, Irvine
ses@ics.uci.edu

Abstract

Requirements Engineering (RE) research is believed to be mature enough for the community to be able to make comparative evaluations of alternative tools, techniques, approaches and methods. Commonly used exemplars in RE that have emerged over the years all suffer from well-defined and widely accepted evaluation criteria which makes comparison of the effectiveness of different research outcomes impossible. The first International Workshop on Comparative Evaluation on Requirements Engineering was held in conjunction with the 11th IEEE International Requirements Engineering Conference in Monterey Bay, California. This workshop was conceived to address these issues and facilitate a community initiative in developing a common understanding of evaluation criteria and developing benchmarks for comparative evaluation in RE. Content, of course, is important.

Keywords: requirements, evaluation, benchmarking

Introduction

The need for an assessment of the progress made in RE research has been felt across the RE community for several years. A number of requirements and specification exemplars [1] have appeared along the years (e.g., the meeting scheduler, the London ambulance computer aided dispatch system, the light control system). These exemplars have been useful for illustrating new RE tools, techniques and methods, and for identifying potential lines of research. However, the commonly used exemplars in RE all lack well-defined *evaluation criteria*, thus making comparison of the effectiveness of the different approaches impossible. Some of the more mature methods and tools in RE have been subjected to pilot studies [2] in industrial settings. While these provide a good indicator of the utility and effectiveness of such methods and tools, they tend to focus on improvements to the technique under study, rather than providing any basis for comparison with alternative techniques.

The first International Workshop on Comparative Evaluation in Requirements Engineering was held on 8th September 2003 in conjunction with the 11th IEEE International Requirements Engineering Conference in Monterey Bay, California. The organisers of the workshop felt that research in RE has become sufficiently mature for the community to begin to make detailed comparative evaluations of alternative techniques. For example, although RE processes are extremely rich and varied, it is possible to identify areas that are sufficiently understood to allow the definition of *benchmarks* [3]. The utility of such benchmarks for both research and industry has been clearly demonstrated by analogous efforts in other fields, e.g., the TREC competition in text recognition and RoboCup (robot soccer) in robotics. By their very nature, successful benchmarks need a community effort to be defined and established. In seeking to define an agreed benchmark, research communities often experience a great leap forward, both in terms of collaboration and consensus among researchers, and in terms of technical results. This workshop was established with the aim of facilitating a community initiative in this direction.

Contributions were solicited in a number of areas including the following:

- Research method and research validation in RE:
  - How do we choose our research goals?
  - How do we evaluate success?
  - How do we measure the impact/importance of a research program?
  - Should we be more explicit about our research methods?
- The role of comparative evaluation in RE:
  - Establishing the necessary consensus on how to compare research results
  - Strengths and weaknesses of various comparative evaluation approaches
  - Experience of these evaluation approaches in other fields
- Determining which sub-areas in RE are ready for comparative evaluation:
  - Identifying task samples and evaluation criteria
  - Proposing potential benchmarks for specific RE activities.
- Reporting on the results of empirical studies and comparative evaluation of RE techniques, methods and tools.

In response to the call, 9 papers were submitted, and were peer-reviewed anonymously by three (for position papers) or four (for full papers) program committee members. Based on the results of the reviews, the program committee selected 6 papers (4 full and 2 position) for presentation at the workshop, and publication in the workshop proceedings.

Structure of Workshop

The workshop was structured to favour discussion and interaction among participants over presentations. Preliminary versions of accepted papers were made available electronically to all participants before the workshop and discussants were appointed for each presented paper.

The morning session was dedicated to the presentation of accepted papers. Discussants were asked to present a response to the paper immediately after the author's presentation. Each paper was allocated 15 minutes for presentation, followed by 5 minutes for the discussants, and 10-15 minutes for further questions and plenary

discussion. Summaries of these discussions are reported in Section 3. The afternoon began with a keynote presentation by Colin Potts, Georgia Institute of Technology, and the rest of the day was dedicated to break-out sessions on specific themes. These were: Benchmarks for RE, Measurement in RE, Experimentation in RE, and Sharing, Not Comparing RE. At the conclusion of the day, summaries of discussions from each breakout session were presented to the attendants (these are also reported in the next section) and the workshop ended with some concluding remarks from the organisers.

Paper Discussions

Session 1: Positions.

This session included two position papers, from Roel J. Wieringa and Kimberly S. Wasson. The first paper by Wieringa provided a much needed reflection about the nature of research in RE. In particular, the paper positioned RE as a *knowledge problem*, i.e. as the problem of increasing the researcher knowledge about the world (as opposed to an *action problem*, where the stated desire is to change the world, which is where design problems live). As such RE research can never result in the *prescription of a method*. The fifteen controversial claims stated in the paper spurred a very lively discussion. A number of votes were called for from the workshop attendees on the most controversial claims from the paper. The results are summarized in Table1

Claim	Question	Agree	Disagree	Not sure
#2	RE isn't really "engineering" as it is not about changing the world (that's the province of *design*)	4	11	7
#15	Benchmarks do not play a role in RE because RE is about problem analysis rather than solution design	4	20	4
#12	RE research cannot produce methods as its outcome	1	30	0
#4	RE research is about building partial theories and RE practice is about building domain theories	7	5	12

There was a general consensus between the attendees that a "purist" vision of RE research, as the one proposed in Wieringa's paper, was too restrictive in light of the social implications that RE research has on the real world. Also, it appears that the majority opinion about the most relevant claim for the purpose of the workshop, number 15, endorses the workshop's goal of raising awareness about the role of evaluation in RE. It is interesting to note that, after almost two decades of specific research in RE, the community is still divided on the fundamentals of the discipline (e.g., what RE and RE research are about).

The paper by Wasson focused on the theoretical and practical difficulties in building and running significant benchmarks in RE, especially when these benchmarks include measures of human behaviour, as is often the case. In these benchmarks, particular attention should be paid to ethical issues, and to differentiate between essential features and environmental influences. Despite all the difficulties, the paper concluded that benchmarking could provide more solid and compelling results than those that we have seen so far in RE. The paper also contributed some initial thoughts on establishing a benchmark to evaluate comprehensibility of requirements specifications.

The participants agreed that the difficulties pointed out by Wasson are substantial, but still results in this area (even if at first unsatisfying) are necessary. Also, comprehensibility was agreed to be one of the core aspects to focus on. The general discussion that followed raised a number of important points. On one hand, proposals from the RE community seem to be finally scaling up to large problems, and concentrating on single aspects for evaluation purposes risks moving the focus out of the large picture and back to minutiae. On the other hand, evaluation of specific aspects would allow people to concentrate on sub-areas that are sufficiently understood, and progressively accumulate small pieces of well-founded evidence. More solid theories and a more thorough understanding of the complex phenomena in RE can then be built on top of this sound evidence. The debate is indeed an old one, as it pits an analytic view of science against a synthetic, or holistic, view.

Session 2: Frameworks.

This session included two papers on specific evaluation frameworks, by Martin S. Feather and Ban Al-Ani.

The first paper by Feather presented the TIMA approach to risk management, supported by a tool called DDP. He argued that evaluation of RE methods and tools can be reformulated as the question "how does the adoption of such methods and tools put my development goals at risk?". In this view, evaluation in RE is not dependent only on the specific artefact that is being evaluated, but also on the goals of a potential user, and – critically – on the strategies that can be used to mitigate the risks induced by their adoption.

The discussion that followed pointed out that in this approach evaluation of RE research is very similar to evaluation of new technology, and that TIMA is actually an interesting technique for general decision making. Indeed, the approach could be applied to TIMA itself (as an evaluation of the DDP tool), to decide whether to use it or not. Unfortunately, due to timing constraints it proved impossible to run an experimental evaluation using TIMA at the workshop itself, as had been proposed in the paper. Several attendants expressed great interest in such an exercise.

The RAV evaluation framework proposed in Ban Al-Ani's paper distinguished three layers for evaluation: the *standards layer*, focusing on recognised standards compliance, the *empirical layer*, collecting empirical methods proven effective in the past, and the *industry layer*, concerning compliance with *de facto* standards adopted by industry. As the framework was still a work in progress, no definitive results on its application were reported. The question was raised about the importance and the role of industry trials of RE research artefacts especially when these artefacts are generally heterogeneous (e.g. tools, methods, techniques, experiments, and understanding of a problem). The framework suggested by Al-Ani was mainly interested in evaluation with the overall purpose of industry uptake of the research artefacts.

Session 3: Experiences.

The last paper session included two papers by Daniela E. Damian and Uolevi Nikula. Damian presented her experiences about running laboratory experiments, with a particular emphasis on the choice of research methodology for this kind of studies. In particular, the methodology described was applied to verify whether the geographical location of stakeholders had any impact on the effectiveness of negotiations during requirements analysis. Damian's experiment provided an example of how designing an evaluation provided insight into the underlying RE problem. Consequently, the lessons learned had implications beyond a single empirical result. A racetrack metaphor was used throughout the discussion. Earlier in the day, benchmarking was described as racing two tools or methods around a racetrack. To extend the analogy, the learning occurs when we come to consensus on what is the racetrack and analysing how cars finish, or fail to finish, the race.

Nikula reported on how the BaRE method he had developed was evaluated by comparison with other established frameworks and by surveying industrial practice in three case studies. Results from the theoretical and from the industrial evaluation were found to be "complementary".

From the discussion that followed, it emerged that this kind of exercise (evaluating a method using another method as a basis) really needs some kind of reference point or benchmark to anchor the chain of evaluations to some agreed-upon, firm basis. The lack of consensus in the community about such a reference point is indeed an obstacle in the path of method evaluation.

Keynote Talk

The second half of the workshop began with a keynote talk by Colin Potts. He began his talk by re-visiting the different arguments that were put forth over the course of the morning. He then situated them within an metaphysics and epistemology of RE, that is, what is the fundamental nature of the problems we are concerned with, what is RE, and what is knowable and achievable in the dominant view. This deep reflection was accompanied by many cartoons, self-deprecation, and other sources of levity.

His talk concluded with three points. The first point was that comparison needed a common basis, that is, it is only fruitful to compare apples to apples and not apples to oranges. To this end, we need to identify meta-problem frames that differ in common bases. Potts suggested that it would be preferable to integrate rather than compete. The second point was that RE is concerned wicked problems. As such, we shouldn't use exemplars that are easily described, such as libraries, elevator scheduling or conference organization, and we shouldn't hide behind the difficulties of instrumentalism. At its heart, RE (and system development) is about social intervention. The third and final point advised use that evaluation should be case-based, because systems are more tangible than methods of working. In this respect, technical artifacts are better lenses than work practices.

Pott's talk provided an insightful and humorous foundation for the breakout sessions that followed.

Breakout Sessions

After the keynote talk, breakout groups to carry forward the discussion in specific areas were formed. The audience proposed a number of possible themes for the breakout groups; these were ranked by popularity and reduced via multiple votes until four reasonably large and general themes were identified. Each group was given 1 hour and half to discuss in greater depth each of the subjects, after which time a presentation was given to the audience, highlighting the major conclusions of the discussion. Summaries of these presentations are given in the following.

Group 1: Benchmarks.

The first breakout group convened with the aim of discussing benchmarks, as a specific form of evaluation, and possibly of defining a specific benchmark for a given area of RE. An attempt was made to classify the *objects* to be benchmarked according to the phase of a typical RE process where the techniques, methods and tools are used. Thus, an initial classification had *elicitation*, *modelling*, *validation*, *communication and documentation* and *evolution* as principal components (or *leagues, to use a term from sports*). Leagues, or pairs of complementary technologies, were to be identified. During the discussion, KAOS and i* were cited as an example pair for the modelling league. Other examples of champions waiting for a matching element to form a pair included the WinWin approach, UML, SCR, RSML and JAD. Also, in the validation league, property proving and error finding were cited as comparable elements (in that an error can be seen as the violation of a desired correctness property).

It was widely acknowledged that a number of factors could interfere with the attainment of the very goals of benchmarking. An ideal benchmark should be reliable, accurate, and complete. It should not make unreasonable assumptions about time, cost, or platform requirements. Also, a good benchmark should be independent of, or explicitly take into account, the structure and practices of the organization where the RE process is being conducted. Of course, these practical constraints (i.e., only a small number of people can be accessible for benchmarking, and these can be not representative of the entire population at large) translates into as many threats to the validity of the results of benchmarking. Finally, given the huge effort required to establish, set up, and run a benchmark, it is not always clear whether the benefits justify the costs. This may explain why industry prefers to simply consider comparing to the "industry best" (i.e., the most successful company on the market) to be a more practical way to satisfy their overall evaluation goals.

The group discussion also focused on the issue of benchmark evolution, and in particular on how benchmarks are introduced (via a community effort or formal standardization process), how they evolve through different revisions, and how they are retired at the end of their life (either because they have been "solved", in that the features measured have become common practice, or because they become too narrow and can no longer serve to distinguish different stages of quality in addressing a specific problem).

The issue of complementing existing exemplars (e.g., the elevator problem) with appropriate measures of success of the solution in order to turn them into benchmarks was also discussed, but no concrete proposal could be made in the available time.

Group 2: Experimentation.

The second breakout session focused on the role and importance

of experimentation in RE. The RE community generally accepts the important role that experimentation can play but is more interested in how to do it well. Experiments could also be conducted for the purpose of comparison or to confirm a theory. The group could not imagine anyone doing an RE experiment to confirm a theory. The high level objectives that were identified for experimentation were: understanding, repeatability and confidence. An important question was raised as whether or not research will be valuable if it does not produce results that will be adopted by practitioners.

If experiment is a study to illuminate a relation between controlled variables and dependant variables, we need to define these variables. The group managed to reach a consensus on dependable variables being cost, time, risk and the like. Much less consensus was there on controlled variables other than "my method versus some other method." There was also some discussion on the boundary conditions of one of the controlled variable: using students versus professionals as the subjects of experiments. The discussion ended with revisiting the question: "how can we make a convincing RE experiment? The general consensus was that we still don't know but this session helped us by developing some new questions to reflect upon.

Group 3: Measurement.

This break out session started out by focusing on three questions: a) What are the assessment factors? b) What should we be measuring?, and c) What challenges are there? Finding satisfactory answers to these fundamental questions seems to be essential. The group then discussed the difference between assessment and evaluation. They were interested in determining a couple of universals in all the RE processes that RE community is familiar with. Examples included *group work*: communication between two or more groups of people, *writing requirements*, *elicitation*: getting knowledge from one head to another. For example, how do we measure if elicitation activity has succeeded. We could run a diagnostic test and see if what is understood could be articulated to a non-domain expert. These are behavioural indicators that require a scientific theory to define and measure and essentially it is difficult general enough indicators. The group felt that we could make a start by going for a rough and ready indicators to measure and operationalise rather than a complete measure.

Another measurement could be performed on the understandability of the requirements document by a member of a specific functional group such as testers. In other words, a really good requirements document must be written so well for testers that they can derive tests straight from it. One measure could be the number of times a tester has to go somewhere else to find information to write tests. This could also be done for architects and designers. One could also use a survey approach to measure testers (architects, or designers) satisfaction with the quality of requirements document. But the group agreed that the ultimate measure of the quality of requirements document is the validity of the end product. The following measures were identified for evaluating the quality of requirements document: a) the number of times testers found in the requirements document ambiguity, incompleteness, and inconsistencies; b) existence of pre- and post-conditions; c) size of requirements specifications; d) navigability; e) time taken to complete; f) Satisfaction in terms of confidence, detection of errors and minimal frustration; g) presence of test criteria (testability)

***Group 4:** Sharing, Not Comparing.*

The fourth group was comprised of participants who where interested in exploring sharing rather than comparative evaluation as a means to achieve a better understanding of RE research. This scepticism was both healthy and welcome.

They felt it was premature for RE to be thinking about benchmarks and it would be more beneficial at this point to work on sharing of research results. Some proposed mechanisms for this cooperation were web sites with collected results, links to tools, and exchanges among researchers. Other possibilities included finding complementary technologies and researchers, as well as identifying commonalities across projects and results. Their ideas were refreshing and could easily be pursued in parallel with comparative evaluation. Moreover, increasing the level of sharing in the community could in turn make us more mature and more prepared to tackle benchmarking.

Conclusion

The workshop was concluded with much support and enthusiasm from the attendees and RE04 organisers for this initiative to continue. The workshop was perceived to have been successful in bringing the important issue of comparative evaluation of RE research effort into highlight. It was agreed that much effort is still needed for the RE community to agree on some of the fundamental issues in RE research. The workshop was perceived to be a right move in that direction.

Potential expected outcomes for future Workshops were identified as follows:

1. A classification scheme for research methods and validation techniques for RE research, together with strengths and weaknesses for each.
2. A consensus-based framework for benchmarking in RE;
3. Specific benchmarks for some basic RE activities;
4. Evaluation criteria to go with well-known exemplars.

Acknowledgements

The workshop organizers gratefully acknowledge the time and effort of members of the Program Committee, RE04 organizers for providing professional facilities and services for running the workshop, the attendees to the workshop for their contribution to discussion and sharing their notes of the workshop for the preparation of this summary, in particular Thomas Alspaugh and Martin Feather.

References

[1] M. S. Feather, S. Fickas, A. Finkelstein, and A. van Lamsweerde, "Requirements and Specification Exemplars," Automated Software Engineering, vol. 4, pp. 419-438, 1997.

[2] R. L Glass "Pilot Studies: What, Why and How" J. Systems and Software, vol 36, no 1, pp85-97, 1997

[3] S. Sim, S. M. Easterbrook and R. C. Holt "Using Benchmarking to Advance Research: A Challenge to Software Engineering". Proceedings of ICSE-2003, pp. 74-83, 2003.