

Comparative Evaluation: Implications from the Multidisciplinary Nature of Requirements

Kimberly S. Wasson

Department of Computer Science
University of Virginia
151 Engineer's Way, P.O. Box 400740
Charlottesville, VA 22904-4740, USA
ksh4q@cs.virginia.edu

Abstract. A challenge has been put forth that the software engineering community organize to advance its scientific maturity through benchmarking. It has been further asserted that requirements engineering is ripe for activity in this area. The position of this paper is that while there are very likely benefits to be gained from movement in this direction, implications from the multidisciplinary nature of requirements warrant very careful consideration if value is in fact to be achieved. The challenges provided are both intellectual and practical. First, the complexity of the various socio-technical goals of requirements complicates the definition of clear and useful but testable performance measures. Second, the fact that the most interesting performance measures involve the behavior of humans places non-trivial constraints not only on what kinds of data can be gathered practically in the service of this goal, but on what kinds of research activities may be undertaken ethically. These issues are illustrated with examples from the author's experience, and a way forward is proposed.

1 Introduction

The software engineering community has been challenged to undertake an initiative toward comparative evaluation, or benchmarking. It is asserted that since other research communities have seen benefits such as stronger consensus about goals and more rigorous examination of research results, that software engineering could benefit from proactive organization toward a culture that includes comparative evaluation [3]. A discussion has been initiated to explore this possibility and its direction, and requirements engineering has been suggested as an area that might be ripe for benchmarking.

The position of this paper is that while some of the benefits that benchmarking has provided to other disciplines might be realizable for requirements engineering, implications from the multidisciplinary nature of requirements warrant very careful consideration in order to preserve this possibility. Further, certain benefits threaten to be prohibitively costly to achieve.

The challenges provided are both intellectual and practical, since the process of constructing a benchmark implies that we must rigorously define both what we want to measure, and how we want to measure it. Such decisions may be arrived at intuitively in some areas, such as heuristics for optimization problems, where, for example, the quality of a solution and the time required to produce it are of low-dimensionality and thus easily defined. Further, the gathering of these measures is easily automated. How-

ever, both defining and collecting meaningful performance measures given the multidisciplinary nature of requirements poses challenges. To meet them, this developing community must consider carefully the issues to be discussed in this paper.

Intellectual issues surrounding the definition of performance measures for requirements are addressed first, with illustration from an experience in empirical work on requirements. This is followed by a discussion of practical issues that further derive from requirements' multidisciplinary nature. Finally, a performance measure for a task that is both difficult and common to virtually every requirements undertaking is proposed that functions within the constraints specified. In this way, it both retains the potential to provide useful and meaningful comparative data, as well as provides an opportunity to explore and assess more fully the ramifications of these issues.

2 Intellectual Issues

Rigorous definition of what to measure is an intellectual concern for comparative evaluation of requirements engineering. It is recognized in [3] that there is a qualitative difference between familiar performance measures and what we are likely to be interested in for software; let us consider this issue further.

Defining what to measure within requirements poses the challenge of navigating very complex socio-technical processes and choosing parts of them that are influential and common enough to realizing the goals of requirements that we are interested in them. The decision of what is meaningful to compare is not straightforward, or easily testable; further, there exists tremendous variety in the elements of, and the entities that influence, requirements processes across organizations and applications.

This has a peculiar effect on the shape of a benchmark in that the interesting objects of measurement are likely to be compound variables, corresponding to emergent properties. Not only must we determine how to operationalize such complex notions, but it could be difficult to tell which of the possible dimensions contributed in what ways to the eventual outcome. This is, however, the nature of the beast; it is argued in [1] that the "messy" problems are the important ones. The road to better understanding is paved with care and rigor.

For example, a ubiquitous performance tradeoff is the speed-quality-cost triad. The common wisdom says that you can pick two. Various existing benchmarks in other areas measure simple instantiations of these properties, for example, the speed at which a processor can process a data set. Processor speed is an intuitively meaningful performance measure that is straightforward to operationalize: the processor is run on the data set and the elapsed (processor, wall clock or other) time is recorded.

How does this translate to requirements? For example, what does "quality" *mean*? It can be argued that a necessary component of requirements quality is the comprehensibility of the documented intent or meaning to a developer who needs to work from those requirements. Quantifying (or even qualifying) comprehensibility is more difficult than quantifying processor speed, since it involves a relationship between a representation and a person, instantiated by hidden psychological activities occurring in that person's mind. In an experiment conducted by the author to determine the impact on potential developers of using a particular method of requirements presentation, com-

prehensibility was operationalized via subjects' performance on a diagnostic test of domain knowledge represented in the experiment materials. While this did allow a contrast between the overall performance of the control and experimental groups to be made evident, the measure did less to discern which particular factors of the experimental presentation allowed the improved performance. Thus when a property is multi-dimensional or emergent, it is more difficult to disentangle the contributing influences.

A successful benchmark for comprehensibility of a requirements presentation would need to go at least this far in defining the performance measure, and ideally further, such that competing influences might be distinguished. In addition, it must consider how the measure might be normalized such that variant methods could apply it; this enables the comparison. With attention to such issues, however, progress can and should be made.

3 Practical Issues

In addition to intellectual issues, the multidisciplinary nature of requirements, and in particular, the fact that the most interesting performance measures involve the behavior of humans, places non-trivial practical constraints on *how* we can measure what we want to measure.

First, the success criteria implied by the design of a benchmark must be amenable to demonstration given available methods and at the level of confidence accepted as standard in the scientific communities upon which requirements engineering draws. In particular, valid data collection mechanisms in the social sciences are strictly defined. For example, there exists a discipline of survey writing that addresses the theory and practice of building the form and content of surveys and questionnaires that generate minimally flawed data. Among the issues this discipline addresses are strengths and weaknesses of various question formats; if human subject data is to be gathered for requirements research via surveys or questionnaires, the objects of measurement must be very carefully coordinated with question formats such that bias or other anomalies have minimal chance of corrupting the data.

In addition to issues of instrumentation, there are issues of coordination of subjects. While a theory group might purchase a set of dedicated machines on which to run optimization experiments, choreographing human subjects is far more difficult. Human subjects not only cannot be dedicated, they require enticement (or at least free time and good will). Convincing a development organization to provide you with access to their most valuable resource is an exercise in finesse and compromise, and such agreements are usually subject to revision that is not in your favor when organizational deadlines loom. Further, many kinds of data collection involving humans cannot be automated; the time required of both experimenters and subjects to gather data sufficient to promote useful analysis can render an experiment prohibitively costly from the point of view of the subjects, the experimenters, or both.

Further considerations affect what data can be gathered ethically; most universities enforce rules of engagement for working with humans that exist to protect the rights of subjects and minimize exposure to risk. The process of gaining approval from

such organizations is often labor-intensive; in our experience it required a person-week to assemble and complete the required materials, followed by a round of revision in accordance with board recommendations. Further, once approval is granted, changes to an experimental protocol are difficult to make and require additional intervention from the review board.

Among the concerns of such boards are the content and application of instruments; in addition to adhering to good scientific practice of data collection as discussed above, benchmark designers must construct instruments that do not unnecessarily expose subjects to risk of, for example, political backlash in the workplace if their survey data is for some reason unattractive to their organization. Thus appropriate confidentiality and other measures must be taken.

Such practical concerns are not insurmountable, but they must be negotiated very carefully in order to preserve the integrity of data and the well-being of subjects. The challenge for a benchmarking initiative is to work within these constraints while defining the measures and formats to additionally be comparable and meaningful across applications of the benchmark.

4 Proposed Target

A subgoal of virtually any requirements undertaking is the intact communication of domain-specific knowledge to a number of individuals who are not experts in the application domain. I propose the undertaking of an initiative to develop a benchmark based on the experiment previously discussed that provides a bounded body of domain knowledge in a form to be determined, accompanied by the task, for users of the benchmark, of making that domain knowledge transparent to non-domain experts. For example, the experiment previously discussed was based on domain knowledge relevant to determining compliance with a standard for maritime track control systems. While the standard was supplied by an industrial collaborator, many such potential target applications are publicly available. In particular, DO-178B, the FAA's standard for compliance of aviation software, is soon to undergo update and either it or its successor could provide a wealth of safety-critical domain-specific information to an undertaking of comparative evaluation.

Evaluation can be accomplished via diagnostic test of the domain knowledge given to non experts, following application of an intervening requirements acquisition strategy. The diagnostic test can be constructed in consultation with domain experts and structured such that grading is consistent and meaningful; further, both closed- and open-form questions can provide a mix of quantitative and qualitative data. Performance measures might include the percentage of knowledge correctly, completely and consistently communicated, the time necessary to complete the test, and the time and effort necessary to apply the intervening strategy. While each of these measures requires further definition in itself, and the test must be constructed, validated and administered within applicable constraints, [2] shows that, for example, comprehension can be operationalized via such a testing strategy.

Along with requirements negotiation, requirements elicitation has been suggested as a core "messy" problem, involving a complex tangle of social and technical factors

[1]. Elicitation relies on the intact communication of domain knowledge. An initiative to comparatively evaluate the fitness of strategies and methods for accomplishing this communication would be of great service to the requirements community.

5 Summary

There are both intellectual and practical concerns for comparative evaluation in requirements that result from its multidisciplinary nature. The definition of clear and useful but testable performance measures is complicated by the complexity of the various socio-technical goals of requirements. Further, the fact that the most interesting performance measures involve the behavior of humans places non-trivial constraints not only on what kinds of data can be gathered practically in the service of this goal, but on what kinds of research activities may be undertaken ethically. In order to gain value from a comparative evaluation initiative, these constraints must be carefully negotiated. Developing successful comparative evaluation initiatives in requirements promises to be far more effortful than in other disciplines, however, this is a necessary step to achieving greater understanding of the most recalcitrant problems we face as a research community.

6 Acknowledgements

It is a pleasure to thank Susan Sim, Steve Easterbrook, and Ric Holt for their ICSE presentation at which this gauntlet was thrown. I also thank John Knight for providing a sounding board for these thoughts. This work was funded in part by NASA under contracts NAG-1-02103 and NAG-1-2290, and by NSF under contract CCR-0205447.

7 References

1. Feather, M., Fickas, S., Finkelstein, A., van Lamsweerde, A.: Requirements & Specification Exemplars. *Automated Software Engineering* 4(4) (1997) 419-438
2. Hanks, K., Knight, J.: Improving Communication of Critical Domain Knowledge in High-Consequence Software Development: An Empirical Study. *Proceedings: 21st International System Safety Conference (2003) to appear*
3. Sim, S., Easterbrook, S., Holt, R.: Using Benchmarking to Advance Research: A Challenge to Software Engineering. *Proceedings: 25th International Conference on Software Engineering (2003) 74-83*