

Requirements Metrics: Scaling Up

Kimberly S. Wasson

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

Abstract. Establishing the relative value of results within a field of study contributes to advancement of that field. In order to compare across large numbers of results together, the methods and metrics used must be scaled up from existing studies where the number of subjects or cases is small. This scaling provides specific challenges, for example, metrics used in small studies are often shaped by factors of the local environment. Standardization is required in order to enable aggregation of data across multiple distributed environments. Further, standardization of metrics is both non-trivial and insufficient. Complex linguistic factors must be accounted for in order to maximize consistency of metric interpretation and use, and a number of other issues must be addressed to ensure that the metrics are interesting as well as practical. This position paper elaborates these issues, sets forth criteria for benchmark-friendly metrics, and proposes a community activity designed to establish a foundational set of requirements benchmark metrics.

1 Introduction

One way to advance the maturity of an area of study is to establish the relative value of contributions within it [4]. Knowing how results compare allows better visualization of progress in the area as a whole, as well as goal-setting based on patterns that become clear through this visualization.

To accomplish such comparison requires a transition from study activities where the number of subjects or cases observed is small to those where they are many. Consider, for example, the goal of determining the utility of methods or techniques for achieving a particular purpose. It is valuable to say that a method or technique has a desired property. It is more valuable to say that one method achieves this desired property better than does another. It is most valuable to say that a number of methods can be completely ordered with respect to a desired property, and to use their distribution with respect to this property to characterize current progress, aid organizational policy decisions, and inform future work. To realize this value, large numbers of methods must be observed under sufficiently comparable circumstances across environments.

The current requirements engineering literature includes numerous studies on one or a small number of subjects or cases (small n), with results that vary in both purpose and quality, and far fewer studies with moderate to large n . RE is an inherently complex area involving both quantitative and qualitative phenomena and drawing from a variety of disciplines; it is not unexpected that there is both variance in the quality of

assessment, as well as difficulty in its effective scaling.

However, effective scaling to larger comparative undertakings is what is needed to achieve the goal of establishing the relative value of contributions. Studies with small n are difficult to aggregate because of variation in locally-applied metrics. Further, purpose-designed benchmarks are impossible unless such variation can be managed. Managing this variation to achieve such scaling, however, is not a simple matter.

The remainder of this paper addresses certain of the challenges that must be overcome in order to achieve the goal of scaling investigations to large n . It further presents a proposal designed to establish and orient a sub-agenda for comparative evaluation research in requirements engineering.

2 Starting Point: Small Studies

The current RE literature contains numerous examples of empirical work that either makes statements regarding the success with which a given method achieves a given outcome, or compares the ability of each of a pair or small group. For example, one study examined the degree to which a particular CASE tool aided in the detection of certain requirements deficiencies, including ambiguities and incompletenesses [6]. In another study, the relative effectiveness of three different prompting techniques for requirements elicitation was compared [2].

Common objects of study are the traditional measures in requirements quality: (un)ambiguity, completeness, consistency, correctness, readability, maintainability, testability, etc. However, actually measuring, for example, completeness, is not straightforward, and researchers must carefully define, and argue the validity of, the phenomenon they call completeness and the indicators of completeness they choose to observe. This process is called operationalization.

In studies with small n , operationalizations are often environment-specific in that some dimension of the property under investigation is either particularly important, accessible to observation, subjective to the analysts, or any combination of the three, based in part on factors of the local environment. For example, in the first study mentioned above, ambiguity and incompleteness are operationalized via indicators called *weak phrases*, phrases argued to “cause uncertainty and leave room for multiple interpretation” [6]. This operationalization is at least partly dependent of local factors insofar as 1) the phrases included in the set appear to have been chosen by the individuals performing the work, without systematic criteria that might predict the same set arising in a different environment with different researchers, and 2) the method of observing the indicator, that is, automatic processing that locates and manipulates weak phrases, makes assumptions about the platforms and formats in use in a given environment.

While operationalizations that are shaped by local factors can serve important purposes in local environments, they are not likely to transfer easily to a broader set of environments, for example, multiple sites seeking to compare their own results, without explicit adjustments and guidance. This does not negate the value of such contributions; local relevance and accessibility to observation are attractive features in an object of study. However, it highlights a fundamental tradeoff in the character of observables when going from small numbers of subjects and cases to larger, distrib-

uted ones. That is, while small n studies can cater to local contingencies in their definition of metrics, working with larger n implies added constraints on what is practicable, given variety across a larger number of environments. This has implications for the design of metrics for use with larger studies that will be discussed in later sections.

3 The Leap to Larger, Distributed Studies

While results from small n studies are valuable to particular (often local) goals, and provide necessary foundations for broader comparisons, there are hurdles that must be overcome in order to achieve such breadth while maintaining the integrity and value of the comparison. The main hurdle to be overcome in order to allow for data aggregation and purpose-designed benchmarks is the standardization of metrics.

There is a fundamental leap from studies where n is small to benchmark undertakings where phenomena must be observed across many environments by multiple different observers. That leap is the necessity that the definition, interpretation, and methods of observation of the phenomena in question be sufficiently consistent across environments that the observations allow true comparison.

However, current empirical work in RE is not yet to this point. There exist numerous definitions in the literature for requirements quality concepts, as a result of environmental specialization, variety of purpose, as well as of researcher preference. For example, the following are all published definitions for *completeness*:

“A complete requirements specification must precisely define all the real world situations that will be encountered and the capability’s responses to them.” [5]

“[S]ituations where a specification entails everything that is known to be ‘true’ in a certain context” [7]

“Are all functions required by the user included?” [1]

“Information is complete when it includes necessary, relevant requirements and/or descriptive material, responses are defined for the range of valid input data, figures are labelled, and terms and units of measure are defined” [3]

Note that these definitions vary in at least two dimensions. First, one focuses on all that is known while the remaining three focus on all that is needed. Second, one focuses mainly on elements in the real world, another focuses mainly on system responses, while the remaining two address both of these aspects. These definitions arise from different sources and have different useful functions, but it would be futile to attempt to aggregate data according to them.

Further, many quality terms are sometimes used interchangeably, or as partial synonyms, or have related but separate lay and technical meanings, making it difficult to know when they refer to the same concept or to different ones. For example, *accuracy* can sometimes be synonymous with *correctness*, and *understandability* shares semantics with *comprehensibility*.

It can be useful to have locally-molded definitions that serve particular purposes.

However, such variety does not allow for aggregation of data for even studies that appear on the surface to investigate the same phenomena, nor does it allow for the following step, purpose-designed benchmarks. Standardization is required. But standardization is both non-trivial and insufficient. First, linguistic issues such as those introduced above make standardization a challenge. The difficulties of working with natural language are well-known, indeed they provide some of the most recalcitrant problems in requirements. But the meta-language of requirements, that is, the language we use to talk *about* the theory and practice of requirements, suffers from the same limitations as the language we write requirements *in*. We have the same needs for accuracy, precision, and accessibility in the lexicon of the requirements meta-language that we have for the language of requirements. In particular, issues of ensuring that parties are concerned with the same concept, having the same semantic core and extent, pose difficulty.

In addition to the linguistic issues of concept definition, there are practical issues that affect how we choose what to measure and how to measure it. Linguistic consistency will not be sufficient if agreed upon usages do not also describe phenomena that are interesting objects of study. In addition, they must be amenable to observation, either directly or through indicators, across a variety of environments, and care must be taken that the necessary generalization that allows this does not dilute their value. Further, if indicators are necessary, care must be taken that they are empirically valid.

Thus, criteria for good RE benchmarking metrics are of a number of types. Their definitions need not only have linguistic support for consistent interpretation, but the phenomena defined must also be interesting, and have repeatable, supportable, agreed operationalizations that are comparable across environments.

4 Advancement Proposal

In order to provide a basis for the development of standard, high-quality metrics, I propose the chartering of a working group to assemble and define a set of high-frequency requirements quality terms, their meanings, and uses such that they meet the criteria for RE benchmarking metrics set forth in section 3. It must be noted that the goal is not to invalidate the wide variety of definitions used gainfully for particular purposes, nor is it to declare that for now and all time that these terms shall mean what the working group says they mean (which would be impossible, given linguistic evolution and other phenomena). Rather, the goal is to facilitate our benchmarking goals specifically, while remaining consistent with the general understandings of what these terms actually do have in common across the RE literature. The product and a guide to its intended use would be disseminated through appropriate publication venues in order to lay the groundwork for future studies that would benefit from it.

A tentative starter set for the requirements quality measures that might be addressed includes the traditional measures of requirements quality: (un)ambiguity, completeness, comprehensibility, consistency, correctness, implementability, maintainability, modifiability, readability, writability, testability, traceability, etc. The set would be adjustable as the group sees fit.

The base set of principles to be adhered to in developing statements and uses of

benchmark-friendly metrics are of the several types discussed in section 3. Linguistically, definitions for these concepts must have integrity of both form and content. Practically, the operationalizations must be universally tractable, that is, observable and measurable across environments. In addition, they must be actually useful to investigate. Methodologically, chosen indicators must be empirically sound.

A work product that gathered such a set of metrics combined with the basis for their development and a guide to their use would provide an invaluable resource for the shepherding of larger, rigorous comparative studies in our area.

5 Summary

To undertake requirements benchmarking activities, metrics that are used in studies with small numbers of subjects or cases must be scaled for use with much larger numbers of these entities across diverse environments. Establishing metrics for use with large n has challenges, and a number of hurdles must be addressed in order to meet them. This paper discussed some of these challenges and set forth criteria to characterize good requirements benchmark metrics, taking into account linguistic, practical and methodological issues. In addition, it proposed a working group to organize and drive the production of a foundational set of requirements benchmark metrics. Such a resource has potential to galvanize and orient larger empirical studies in requirements engineering and to make the aggregation and analysis of broad data sets possible.

6 Acknowledgements

I thank the anonymous reviewers for their valuable suggestions. This work was funded by NSF under contract CCR-0205447 and by NASA under contract NAG-1-02103.

7 References

1. Antón, A.: Requirements Engineering. Lecture Slides, http://ecommerce.ncsu.edu/studio/lectures/RElecture_LW.pdf (current on 9 July, 2004).
2. Browne, G., Rogich, M.: An Empirical Investigation of User Requirements Elicitation: Comparing the Effectiveness of Prompting Techniques. *Journal of Management Information Systems*, 17:4 (2001).
3. Radio Technical Commission for Aeronautics: Software Considerations in Airborne Systems and Equipment Certification (DO-178B). Advisory Circular (1992).
4. Sim, S., Easterbrook, S., Holt, R.: Using Benchmarking to Advance Research: A Challenge to Software Engineering. Proceedings: 25th International Conference on Software Engineering (2003).
5. Stokes, D.: Requirements Analysis. *Computer Weekly Software Engineer's Reference Book* (1991).
6. Wilson, W., Rosenberg, L., Hyatt, L.: Automated Quality Analysis of Natural Language Requirement Specifications. Proceedings: Pacific Northwest Software Quality Conference, (1996).
7. Zowghi, D., Gervasi, V.: The Three Cs of Requirements: Consistency, Completeness, and Correctness. Proceedings: 8th International Workshop on Requirements Engineering: Foundations for Software Quality (2002).